Speak in Context: Multilingual ASR with Speech–Context Alignment via Contrastive Learning

Yuchen Zhang[1,2], Haralambos Mouratidis[1,2], Ravi Shekhar[1,2]
[1]Institute for Analytics and Data Science, University of Essex
[2]School of Computer Science and Electronic Engineering, University of Essex
{yuchen.zhang,r.shekhar,h.mouratidis}@essex.ac.uk

Abstract

Automatic speech recognition (ASR) has benefited from advances in pretrained speech and language models, yet most systems remain constrained to monolingual settings and short, isolated utterances. While recent efforts in context-aware ASR show promise, two key challenges persist: limited multilingual support and the absence of principled alignment between speech and contextual representations. In this paper, we introduce a context-aware multilingual ASR framework that supports diverse languages and accents while preserving the modularity of pretrained models. Our approach combines a frozen speech encoder and a decoder-only language model via a lightweight projection module, allowing structured context prompts, including dialogue history and biasing words, to guide transcription. To improve interaction between speech and context, we employ a contrastive learning objective that aligns their representations in a shared embedding space. Evaluations on over 1,500 hours of real-world conversational speech across 11 languages and 5 English dialects show that contextual input consistently improves recognition quality. Contrastive alignment provides additional gains when applied to different context types, with an overall performance gain of over 5%. These results highlight the importance of both contextual modeling and cross-modal alignment in multilingual ASR.

Keywords: Automatic Speech Recognition, Multilingual ASR, Context Information, SpeechLLM, Contrastive Learning

1. Introduction

Automatic speech recognition (ASR) has advanced rapidly in recent years, largely due to the development of large-scale pretrained models and end-to-end architectures. However, real-world ASR systems still face persistent challenges in multilingual scenarios. Recent developments in speech-language model integration have paved new ways to connect pretrained speech encoders with large language models (LLMs), enabling speech-to-text transcription via prompt-based mechanisms while often keeping backbone components frozen (Verdini et al., 2024). For example, Hono et al. (2024) proposes integrating a pre-trained speech representation model with an LLM through a bridge network, achieving competitive end-to-end ASR performance. Fathullah et al. (2024) demonstrates that attaching a small audio encoder to a frozen LLM allows multilingual speech recognition, even though the LLM was trained primarily on English text.

However, two critical gaps remain. First, multilingual context-aware ASR, which supports multiple languages and actively integrates preceding conversational context or biasing lists, has been relatively understudied. For instance, Cheng (2024) investigates prompt-based context-aware recognition in accented speech but focuses on monolingual short-utterance scenarios rather than fully multilingual conversational settings. Second, while many works incorporate context as additional input such as concatenating previous utterances or bias lists (Züfle and Niehues, 2024; Yang et al., 2024) , the explicit alignment between speech embeddings and contextual embeddings via a trainable, embedding-level modality alignment mechanism remains largely unexplored. For example, Guo et al. (2021) introduces a context-aware language model that encodes callsign lists alongside ASR decoding but does not explore embedding-space alignment between acoustic and context representations.

To address these gaps, we propose a context-aware multilingual ASR framework [1] that supports cross-lingual recognition while explicitly aligning speech and contextual representations in the embedding space, moving beyond heuristic concatenation. Our method integrates a frozen speech encoder with a frozen decoder-only LLM through a lightweight projection module, and incorporates structured context information into the LLM input. To enhance cross-modal alignment, we introduce a contrastive learning objective that draws speech–context pairs closer in the shared representation space. This design enables the model to condition generation on both acoustic and contextual cues without modifying the underlying pretrained components.

We conduct extensive experiments on the official dataset for the Interspeech2025 **M**ulti**L**ingual **C**onversational **S**peech **L**anguage **M**odels (MLC-SLM) challenge (Mu et al., 2025), a large-scale

[1]The code is available at https://github.com/yuchen-zhang-essex/Context-Aware_ASR.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5873–5882
11–16 May 2026.

multilingual conversational dataset spanning 11 languages and over 1,500 hours of real-world speech. The experimental results suggest that incorporating context consistently enhances transcription quality, validating the benefits of context-aware generation across diverse linguistic conditions. Our findings further highlight the importance of aligning contextual and acoustic representations, showing that contrastive learning offers additional improvements on different types of context. These results underscore the need for more principled speech–context integration approaches in multilingual ASR.

Our main contributions are as follows:

- We introduce a context-aware SpeechLLM framework for multilingual ASR that effectively harnesses contextual inputs, including dialogue history and biasing words, to enable efficient adaptation across diverse languages, while maintaining a lightweight design.
- We propose an embedding-level speech context alignment strategy based on contrastive learning, explicitly linking speech features with contextual information to improve semantic grounding in multilingual scenarios.
- We conduct comprehensive experiments on a 1,500-hour multilingual dataset across various context settings. Results show consistent improvements over non-contextual and non-contrastive settings, achieving over a 5% overall performance gain, and provide insights into how contrastive alignment interacts with different context types in multilingual settings.

2. Related Work

Multilingual ASR Recent work in ASR has increasingly focused on multilingual settings, where a single model needs to support many languages, dialects, or accents. For example, Babu et al. (2021) introduces XLS-R, a self-supervised model trained on nearly half a million hours of speech across 128 languages, based on the wav2vec2 (Baevski et al., 2020). Their evaluation covers both high- and low-resource languages, achieving large error-rate reductions relative to prior monolingual models. Similarly, Khurana et al. (2022) propose SAMU-XLSR, an utterance-level multimodal multilingual speech representation model that aligns speech and text embedding spaces across languages by combining XLS-R and multilingual text embeddings. Other multilingual ASR works extend the unified modeling across many languages, including low-resource ones (Li et al., 2025). These efforts highlight the value of shared multilingual representation and backbone models for multilingual ASR tasks.

Context-aware ASR Injecting external context, such as preceding utterances, bias-lists of rare words, or domain-specific vocabulary, has been shown to improve ASR performance in particular scenarios, such as rare words, proper nouns, and conversational settings (Concina et al., 2025; Linke et al., 2025). For instance, Chang et al. (2021) propose Context-Aware Transformer Transducer, a Transformer-Transducer architecture that attends over preceding utterances encoded via pretrained BERT or BiLSTM, enabling multi-turn dialogue context to influence recognition. Huang et al. (2020) introduces Class-LM & Word Mapping for contextual biasing in end-to-end ASR, allowing a beam search to traverse into a context FST comprised of rare or domain-specific vocabulary, thus improving recognition of named entities. Beyond these, Fu et al. (2023) presents Robust Acoustic and Semantic Contextual Biasing, where attention-based modules incorporate both acoustic and semantic context cues for rare words in neural transducers. Gong et al. (2024) discusses Contextual Biasing Speech Recognition in dynamic settings, injecting contexts into earlier encoder layers and assessing runtime cost.

Speech LLM With the proliferation of LLMs, recent research has started to bridge speech encoders and LLMs for tasks such as ASR, speech translation, or spoken language understanding. For example, Fan et al. (2025) proposes AlignFormer, a neural adapter connecting a frozen speech encoder to a frozen instruction-following LLM. Their method uses CTC and dynamic-window QFormer layers to align heterogeneous modality lengths and preserve the LLM's instruction-following capabilities. Chen et al. (2024) present SALM, which integrates a frozen LLM, audio encoder, and LoRA layers for speech recognition and speech translation with an in-context learning capability. Hono et al. (2024) proposes a bridge network that maps speech encoder outputs into the LLM embedding space, compressing the sequence length for efficiency. Other work explores end-to-end architectures combining pretrained speech encoders with LLMs for ASR (Luu and Bojar, 2025). These contributions demonstrate the viability of freezing large backbones and training lightweight modules to connect modalities for ASR.

3. Methodology

3.1. Problem Formulation

Given a spoken utterance s_t^j from dialogue j at turn t, along with context information $\mathcal{P}_{ctx_t^j}$, the objective is to generate a textual transcription $y_t^j = \{y_1, y_2, \ldots, y_L\}$, where L is the output length, that

corresponds to the spoken content, considering both the audio signal and the available contextual information.

Formally, we aim to model the conditional probability distribution:

$$P(y_t^j \mid s_t^j, \mathcal{P}_{ctx_t^j}) = \prod_{l=1}^{L} P(y_l \mid y_{<l}, s_t^j, \mathcal{P}_{ctx_t^j}), \quad (1)$$

where y_l denotes the l-th token in the output sequence, and $y_{<l}$ represents all previously generated tokens.

3.2. Overall Structure

The overall structure of the proposed context-aware multilingual ASR system is illustrated in Figure 1. Our model architecture integrates a frozen speech encoder, a frozen decoder-only LLM, and a lightweight projection module that bridges the two modalities. To support context-aware transcription, we first construct a model input by injecting contextual information into a fixed instruction template. This prompt and the speech embedding are passed to the LLM, which autoregressively generates the transcription.

The overall process consists of three main stages: (1) extracting contextual information relevant to the input speech (Section 3.3); (2) projecting the high-dimensional speech features into the shared embedding space of the LLM and aligning them with the contextual prompt. This is achieved through a trainable speech connector (Section 3.4) and a contrastive learning objective (Section 3.5) that encourages semantically related speech–context pairs to be close in the representation space; (3) generating the final transcription conditioned on both the speech input and the injected context using the LLM decoder.

3.3. Context Extraction

In this work, we focus on two types of contextual information to support multilingual ASR: dialogue history and biasing words. Dialogue history refers to the preceding utterances within the same conversation and serves to ground the interpretation of the current utterance in the prior context. This is particularly beneficial for resolving incomplete phrases, pronouns, and context-dependent expressions whose meaning depends on earlier dialogue turns.

In contrast, biasing words refer to keywords or phrases provided in advance, such as named entities, domain-specific terminology, or task-relevant keywords. These terms may be rare or unseen in the training data, but are essential for correctly recognizing domain-relevant content. By jointly incorporating dynamic context (dialogue history) and static prior knowledge (biasing words), we aim to improve the model's capacity to process complex multilingual speech with greater accuracy and robustness.

3.3.1. Dialogue History

To incorporate contextual information, we define the dialogue history of each utterance as the sequence of preceding turns within the same dialogue.

Let the dataset $\mathcal{D}$ consist of a set of M dialogues: $\mathcal{D} = \{\, \mathcal{D}_i \mid i = 1, 2, \ldots, M \,\}$, where each dialogue $\mathcal{D}_j = \{\, s_t^j \mid t = 1, 2, \ldots, N_j \,\} (j \in M)$ consists of an ordered sequence of N_j pieces of speech utterances (N_j may vary across dialogues). s_t^j denotes the t-th utterance in $\mathcal{D}_j$ and its corresponding ground-truth transcription is denoted as y_t^j.

For a given utterance s_t^j in dialogue $\mathcal{D}_j$, we define its dialogue history window of size K_{DH} as $\mathcal{DH}_t^j$:

$$\mathcal{DH}_t^j = \begin{cases} \{y_{t-K_{DH}}^j, \ldots, y_{t-1}^j\}, & if \quad t > K_{DH} \\ \{y_1^j, \ldots, y_{t-1}^j\}, & if \quad 1 < t \leq K_{DH} \\ \varnothing, & if \quad t = 1 \end{cases}. \quad (2)$$

To integrate the extracted dialogue history into the model input, we convert $\mathcal{DH}_t^j$ into a natural language prompt. When $\mathcal{DH}_t^j \neq \varnothing$, we construct the Dialogue History Prompt $\mathcal{P}_{DH_t^j}$ as:

"`The previous` $|\mathcal{DH}_t^j|$ `turn(s) of this speech is:` $\mathcal{DH}_t^j$."

Here $|\mathcal{DH}_t^j|$ denotes the number of previous utterances available (up to K_{DH}), and the entries in $\mathcal{DH}_t^j$ are concatenated with a separator token [SEP].

If no conversation history is available (i.e., t = 1), the Dialogue History Prompt $\mathcal{P}_{DH_t^j}$ is defined as :

"`There is no conversation history of this speech.`"

It should be noted that during training, the dialogue history of each utterance is extracted from the ground-truth transcriptions. During inference, since ground-truth transcripts are unavailable, we rely on coarse transcriptions generated by a pre-trained multilingual Connectionist Temporal Classification (CTC) model (Lugosch et al., 2022). The same history extraction and prompt formatting procedure used during training is applied to these transcriptions to construct the dialogue history for the test data.

3.3.2. Biasing Words

To improve contextual awareness and provide the model with additional lexical cues, we incorporate biasing words, which consist of two categories: (1) Hotwords, short n-gram phrases extracted directly from transcriptions, and (2) Distractor Terms, rare

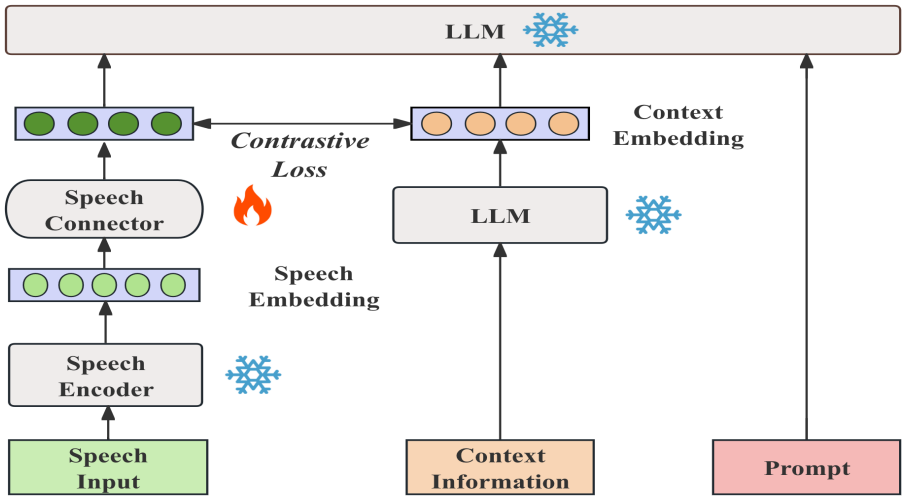

Figure 1: The overall structure of the proposed context-aware multilingual ASR.

words sampled from a predefined lexicon. These words are included as part of the decoding prompt to highlight potentially informative or underrepresented content, while also encouraging the model to develop robustness against irrelevant lexical cues.

The training hotwords are randomly selected from the transcriptions of speech, following a strategy similar to that used in prior studies (Pundak et al., 2018; Huang et al., 2023; Futami et al., 2024; Yang et al., 2024). For a given utterance s_t^j, let its ground-truth transcription be denoted by y_t^j. To construct the corresponding hotword set $\mathcal{HW}_t^j$, we first tokenize y_t^j into a sequence of words. From this sequence, we randomly sample several word-level n-gram phrases. The number of phrases to sample is drawn uniformly from $[1, K_{HW}]$, while the length of each phrase is drawn uniformly from $[1, L_{HW}]$, where K_{HW} and L_{HW} denote the maximum number of phrases and the maximum phrase length, respectively.

Additionally, for each utterance s_t^j, we include K_{DT} distractor terms $\mathcal{DT}_t^j$, sampled from a predefined, language-specific rare-word lexicon. The rare-word lexicon is constructed following a method similar to that of Le et al. (2021). Specifically, we construct the rare word lexicon for each language by first aggregating all ground-truth transcriptions in the training set. Unigram frequencies are computed over the tokenized corpus, and words occurring fewer than a threshold θ_{rare} times are discarded to reduce noise from typos or annotation errors. From the remaining set of words, the bottom p_{rare} fraction, ranked by unigram frequency, is selected to construct the rare-word lexicon. For each utterance s_t^j, the distractor terms $\mathcal{DT}_t^j$ are then sampled from the built rare-word lexicon under the constraint that they do not appear in the current transcription y_t^j,

The final biasing words for utterance s_t^j is formed by combining the hotwords $\mathcal{HW}_t^j$ and distractor terms $\mathcal{DT}_t^j$. We then convert the biasing words into a Biasing Words Prompt $\mathcal{P}_{BW_t^j}$:

"`The speech might contain following words:` $\mathcal{HW}_t^j, \mathcal{DT}_t^j$."

Similar to dialogue history extraction, the source of hotwords differs between training and inference. During training, hotwords are derived directly from the ground-truth transcriptions. Since ground-truth is unavailable during inference, hotwords are extracted from the same coarse transcriptions used for dialogue history extraction. The same sampling procedure and prompt formatting are applied in both cases to ensure consistency across training and inference.

3.4. Speech Connector

Given an input speech s_t^j , the audio encoder extracts a sequence of high-dimensional acoustic embeddings:

$$\mathbf{H}_{spe_t^j}^{raw} = \mathcal{E}(s_t^j) \in \mathbb{R}^{B \times T \times E_a}, \tag{3}$$

where $\mathcal{E}$ denotes the audio encoder, and B, T, and E_a denote the batch size, the number of acoustic frames, and the audio encoder hidden dimension, respectively.

To align the encoder outputs with the embedding space of the LLM, the speech representations $\mathbf{H}_{spe_t^j}^{raw}$ are first downsampled by a factor K_{down}, where every K_{down} consecutive frames are concatenated into a single vector. The stacked features are then transformed through two successive linear layers to get the final representation of speech

$\mathbf{H}_{spe_t^j}$:

$$\mathbf{H}_{spe_t^j} = \mathcal{L}_2\Big(\sigma_{\text{GELU}}\big(\mathcal{L}_1(\mathbf{H}^{stacked}_{spe_t^j})\big)\Big), \tag{4}$$

$$\mathbf{H}^{stacked}_{spe_t^j} = \mathcal{D}_K(\mathbf{H}^{raw}_{spe_t^j}) \in \mathbb{R}^{B\times \frac{T}{K_{down}}\times(E_a\cdot K_{down})}, \tag{5}$$

where $\mathcal{D}_K(\cdot)$ denotes the downsampling operator, $\mathcal{L}_1$ and $\mathcal{L}_2$ denote linear projector, and $\sigma_{\text{GELU}}(\cdot)$ is the Gaussian Error Linear Unit (GELU) activation function.

3.5. Speech–Context Alignment

To enhance alignment between speech and its associated contextual information, we introduce a contrastive learning objective that encourages paired speech and context embeddings to be close in representation space, while pushing apart mismatched pairs.

For a given utterance s_t^j, let the contextual prompt $\mathcal{P}_{ctx_t^j}$ (e.g., dialogue history or biasing words) be tokenized and passed through the input embedding layer of the frozen LLM, denoted by f_{emb}. The corresponding context embedding $\mathbf{H}_{ctx_t^j} \in \mathbb{R}^{E_t}$ is computed as:

$$\mathbf{H}_{ctx_t^j} = f_{emb}(\mathcal{P}_{ctx_t^j}) \in \mathbb{R}^{L\times E_t}, \tag{6}$$

where L is the token length of the context prompt and E_t is the LLM embedding dimension.

To apply contrastive learning with context embedding and speech embedding, we first conduct mean pooling and then L2-normalization to the context embeding:

$$\tilde{\mathbf{H}}_{ctx_t^j} = \psi\left(\phi\left(\mathbf{H}_{ctx_t^j}\right)\right), \tag{7}$$

where $\phi(\cdot)$ denotes mean pooling and $\psi(\cdot)$ denotes the L2-normalization.

Similarly, the projected speech embedding $\mathbf{H}_{\text{proj}}$, which is also aggregated and normalized:

$$\tilde{\mathbf{H}}_{spe_t^j} = \psi\left(\phi\left(\mathbf{H}_{spe_t^j}\right)\right). \tag{8}$$

For each training batch of size B, we define positive pairs as a speech utterance and its corresponding context, and negative pairs as the same speech embedding paired with non-matching contexts from other utterances in the batch. We then compute the pairwise similarity matrix $\mathbf{S} \in \mathbb{R}^{B\times B}$ using scaled dot products between the normalized speech and context embeddings:

$$S_{t,q} = \frac{\tilde{\mathbf{H}}_{spe_t^j} \cdot \tilde{\mathbf{H}}_{ctx_q^k}}{\tau}, \tag{9}$$

where s_t^j and s_q^k denote two utterances in the batch, $\tilde{\mathbf{H}}_{spe_t^j}$ and $\tilde{\mathbf{H}}_{ctx_q^k}$ are their corresponding normalized speech and context embeddings, and τ is a temperature scaling hyperparameter.

The InfoNCE contrastive loss is then computed as:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{B}\sum_{q=1}^{B}\log\frac{\exp(S_{t,t})}{\sum_{q=1}^{B}\exp(S_{t,q})}, \tag{10}$$

where each term compares the similarity between the speech embedding $\tilde{\mathbf{H}}_{spe_t^j}$ and its corresponding context embedding $\tilde{\mathbf{H}}_{ctx_t^j}$ (i.e., the positive pair) against the similarities between that same speech embedding and all other context embeddings in the batch $\tilde{\mathbf{H}}_{ctx_q^k}$(i.e., the in-batch negatives). This encourages the model to assign higher similarity scores to matched speech–context pairs and lower scores to mismatched ones, thereby learning more discriminative and contextually grounded representations.

3.6. Training Objective

For a given utterance s_t^j, the decoder outputs a sequence of logits $\mathbf{Z}_t^j \in \mathbb{R}^{L\times V}$, where L is the output length and V is the vocabulary size. Let the ground-truth transcription be $y_t^j = \{y_l\}_{l=1}^{L}, \quad y_l \in [V]$, where $[V] = \{1, 2, \ldots, V\}$ denotes the index set of the vocabulary, and each y_l is the index of the correct token in this set.

At each position l, the probability assigned to the ground-truth index y_l is calculated by the softmax:

$$p(y_l \mid y_{<l}, s_t^j) = \frac{\exp\big(\mathbf{Z}_t^j[l, y_l]\big)}{\sum_{v=1}^{V}\exp\big(\mathbf{Z}_t^j[l, v]\big)}, \tag{11}$$

where $\mathbf{Z}_t^j[l, v]$ is the logit for vocabulary index v at position l.

The CE loss is then computed as the average negative log-likelihood over all positions:

$$\mathcal{L}_{CE} = -\frac{1}{L}\sum_{l=1}^{L}\log p(y_l \mid y_{<l}, s_t^j). \tag{12}$$

Finally, the total loss is computed as a weighted combination of CE and CL objectives:

$$\mathcal{L} = \beta\cdot\mathcal{L}_{CE} + \alpha\cdot\mathcal{L}_{CL}, \tag{13}$$

where, β is a fixed hyperparameter, and α is dynamically adjusted to balance the two losses:

$$\alpha = \frac{\mathcal{L}_{CL}}{\mathcal{L}_{CE} + \mathcal{L}_{CL}}. \tag{14}$$

The combined objective encourages the decoder to generate accurate transcriptions while enforcing consistency between the speech and contextual representations in the embedding space.

4. Experiment

4.1. Dataset

We use the MLC-SLM dataset (Mu et al., 2025) for the ASR task to demonstrate the effectiveness of our proposed model. The selected dataset contains approximately 1571 hours of multilingual conversational data in total, including 1,507 hours for training (Train), 32 hours for validation (Val), and 32 hours for testing (Test). Each subset comprises 11 languages: English, French, German, Italian, Portuguese, Spanish, Japanese, Korean, Russian, Thai, and Vietnamese. The English subsets contain accents from various regions: American, Australian, British, Filipino, and Indian. Each recording contains a multi-turn conversational speech of around 20 minutes between two speakers on a randomly assigned topic, including celebrities, dreams, education, emotion, fashion, food, games, the Internet, movies, shopping, travel, etc.

Table 1: Durations (hours) of the MLC-SLM dataset.

Language	Train	Val	Test
English-American	100.60	2.22	2.01
English-Australian	100.39	2.34	2.43
English-British	100.48	2.23	2.03
English-Filipino	100.36	2.09	2.02
English-Indian	100.45	2.17	2.31
French	100.38	2.26	2.07
German	100.58	2.03	2.05
Italian	100.67	2.10	2.19
Japanese	100.44	2.08	2.19
Korean	100.68	2.03	2.02
Portuguese	100.33	2.18	2.14
Russian	100.41	2.05	2.18
Spanish	100.47	2.14	2.18
Thai	100.50	2.12	2.17
Vietnamese	100.45	2.15	2.01
Total	1507.22	32.18	32.19

4.2. Experimental Setting

General Setting In this study, we employ Whisper-large-v3 Turbo (Radford et al., 2022) as the speech encoder and adopt EuroLLM-1.7B-Instruct (Martins et al., 2025) as a LLM decoder.

During training, the data input is constructed using a structured textual template designed to guide the LLM. Each input sequence takes the form:

`"<SPEECH> USER: <PROMPT> ASSISTANT: <TRANSCRIPTION>."`

Here, `<SPEECH>` denotes the speech embedding generated by the speech encoder. `<TRANSCRIPTION>` is the ground-truth transcription of the speech. `<PROMPT>` is prefixed task related prompt:

`"Transcribe the speech to text. The following context information might help:<CONTEXT>"`

In our experiments, we define the `<CONTEXT>` in three configurations. When only dialogue history is used, `<CONTEXT>` is set to $\mathcal{P}_{DH}$. When only biasing words are provided, it becomes $\mathcal{P}_{BW}$. When incorporating both dialogue history and biasing words, the context is constructed by concatenating the two: $\mathcal{P}_{DH} + \mathcal{P}_{BW}$.

We also examine the situation where context information is unavailable. In this setting, the <PROMPT> is set as: `"Transcribe the speech to text."`

During the inference process, the format of the data input for the LLM is `"<SPEECH> USER: <PROMPT> ASSISTANT:"`.

Model Configuration The parameters of both the speech encoder and the LLM are frozen. Optimization is applied only to the lightweight projection module responsible for aligning the speech and language modalities. The model is trained for 2 epochs with a batch size of 8 using the AdamW optimizer, a learning rate of 1e-4, and a weight decay of 1e-6. A linear warm-up schedule with 1,000 warm-up steps is employed to gradually increase the learning rate at the early stage of training. During inference, beam search is used for decoding, with the beam size set to 2.

For context-related hyperparameters, the dialogue history window size K_{DH} is set to 1. The maximum number of hotwords per utterance K_{HW} is 3, and the maximum token length per hotword L_{HW} is 3. For each utterance, K_{DT} = 1 distractor term is introduced. A frequency threshold θ_{rare} = 2 is used to discard uncommon terms, and the bottom p_{rare} = 10% of tokens are selected to form the rare-word lexicon. The speech features are downsampled by a factor of K_{down} = 4. The temperature scaling parameter τ for contrastive learning is set to 0.07.

Evaluation Metrics We evaluate model performance using two standard metrics in ASR: Word Error Rate (WER) and Character Error Rate (CER).

WER and CER are computed at the token level by aligning the predicted transcription with the ground truth using minimum edit distance. The error rate is calculated as:

$$ErrorRate = \frac{S + D + I}{N}, \qquad (15)$$

where S is the number of substitutions, D is deletions, I is insertions, and N is the number of tokens (for WER) or characters (for CER) in the reference.

In our multilingual setup, we adopt CER for languages that lack clear word boundaries, including

Table 2: WER/CER (%) results across different contexts (lower is better). Values in brackets indicate the change compared to the corresponding non-CL setting. All = Dialogue History + Biasing Words. BW: Biasing Words; CL: Contrastive Learning. The best results in **Bold** and the second best Underlined

language	No Context	History	BW	All	History+CL	BW+CL	All+CL	metric
English-American	13.29	9.21	10.14	9.42	9.70 (+0.49)	**9.12** (-1.02)	10.05 (+0.63)	WER
English-Australian	12.66	7.93	8.73	8.12	**7.39** (-0.54)	7.78 (-0.95)	7.77 (-0.35)	WER
English-British	8.58	7.46	5.96	5.78	**5.69** (-1.77)	5.85 (-0.11)	6.22 (+0.44)	WER
English-Filipino	10.74	9.27	10.49	11.56	**9.15** (-0.12)	11.69 (+1.20)	9.88 (-1.68)	WER
English-Indian	15.91	9.97	8.43	**8.18**	9.59 (-0.38)	8.92 (+0.49)	8.39 (+0.21)	WER
French	23.32	24.06	27.00	25.01	**22.50** (-1.56)	23.79 (-3.21)	26.10 (+1.09)	WER
German	31.49	19.89	26.14	21.99	**19.36** (-0.53)	20.32 (-5.82)	20.12 (-1.87)	WER
Italian	20.52	25.69	21.78	20.65	20.88 (-4.81)	20.04 (-1.74)	**19.87** (-0.78)	WER
Japanese	38.45	25.65	**19.43**	20.96	21.27 (-4.38)	21.74 (+2.31)	20.60 (-0.36)	CER
Korean	18.15	8.91	7.67	7.73	**7.41** (-1.50)	8.24 (+0.57)	7.74 (+0.01)	CER
Portuguese	44.27	32.09	32.78	31.32	36.66 (+4.57)	35.12 (+2.34)	**29.61** (-1.71)	WER
Russian	**16.88**	20.16	19.69	20.65	18.45 (-1.71)	19.50 (-0.19)	18.90 (-1.75)	WER
Spanish	12.62	12.39	13.79	13.34	**10.33** (-2.06)	11.42 (-2.37)	11.28 (-2.06)	WER
Thai	22.73	22.45	22.58	22.78	**20.50** (-1.95)	23.32 (+0.74)	23.23 (+0.45)	CER
Vietnamese	25.84	12.50	**12.26**	13.69	12.35 (-0.15)	13.78 (+1.52)	13.80 (+0.11)	WER
Avg.	21.03	16.58	16.52	16.08	**15.42** (-1.16)	16.04 (-0.48)	15.57 (-0.51)	-

Japanese, Korean, and Thai. For all other languages, where words are separated by spaces, we report WER for evaluation.

4.3. Results and Discussion

This section evaluates the proposed context-aware multilingual ASR model across 15 languages under different contextual configurations. Table 2 presents the experimental results across different contextual configurations.

Effect of contextual information. Table 2 shows that contextual information consistently improves recognition compared to the no-context baseline. The average error rate drops from 21.03% to 16.08% when both dialogue history and biasing words are provided, while each individual context type also yields notable improvements. These results confirm that contextual grounding significantly enhances multilingual ASR performance. Gains are particularly strong in German, Korean, and Portuguese, though the impact of context type varies. For German, dialogue history leads to the largest improvement, reducing WER from 31.49% to 19.89%, while biasing words are less effective. In contrast, Korean benefits more from biasing words, with CER dropping from 18.15% to 7.67%, compared to 8.91% with history. Portuguese shows gains across all context settings, with the lowest WER of 29.61% achieved when both history and biasing are used with contrastive learning. These results demonstrate the value of leveraging both dynamic dialogue history and static lexical cues, although the effectiveness varies across languages.

Role of contrastive learning. The impact of contrastive learning shows a consistent positive trend across all context settings, though the degree of improvement varies. The best-performing setting is dialogue history combined with contrastive learning, which achieves the lowest average error rate of 15.42%, improving over the history-only setting at 16.58% by 1.16%. This indicates that contrastive alignment is especially effective in leveraging conversational history, helping the model maintain semantic coherence and resolve context-dependent expressions more reliably. When applied to biasing words alone, contrastive learning also yields gains, with the error rate reduced from 16.52% to 16.04%. This suggests that aligning speech with lexical context offers some benefits for recognizing rare or domain-specific terms, but further design improvements are needed to fully exploit its potential. The setting combining both dialogue history and biasing words with contrastive learning produces the second-best performance at 15.57%. Although this improves upon the non-contrastive setting at 16.08 %, the gain is smaller than that achieved with history alone. This pattern suggests that merging heterogeneous context types under a single alignment objective may introduce competing signals. While dialogue history emphasizes semantic continuity, biasing words highlight local lexical anchors, and contrastive learning may struggle to reconcile both simultaneously. Overall, contrastive learning consistently enhances performance across all settings, with the most significant improvements observed when applied to dialogue history alone. These findings underscore the value of targeted contrastive alignment and motivate future work on context-specific or disentangled optimization strategies.

Language-specific behaviors. The results also reveal distinct language-specific patterns. For the English dialects, contextual information consistently improves recognition, with British English achieving the lowest error rate when dialogue history is combined with contrastive learning. French exhibits only modest improvements. While dialogue history with contrastive learning slightly outperforms the baseline, configurations that rely on biasing words often reduce accuracy, indicating that errors in coarse transcriptions or distractor terms can introduce misleading signals. For Italian, dialogue history alone degrades performance, but biasing words with contrastive learning provide small yet consistent gains. Japanese achieves its best performance with biasing words alone; adding contrastive learning degrades results slightly, possibly due to challenges in constructing reliable contrastive negatives for its complex writing system. Portuguese benefits most from combining dialogue history and biasing words with contrastive learning, contrary to expectations of degradation, suggesting alignment enhances contextual understanding. Spanish shows strong improvements when dialogue history is paired with contrastive learning. Russian, however, sees limited gains or mild degradation from contextual cues, indicating that cross-turn prompts are less effective under current settings.

Unseen Languages in Pretraining. Thai and Vietnamese, which are not included in EuroLLM pre-training, provide insight intothe challenges of generalization to unseen languages. Vietnamese shows clear gains from contextual information, with WER dropping by roughly half compared to no context setting. Incorporating dialogue history with contrastive learning further stabilizes performance, indicating that cross-turn grounding can transfer effectively even when the target language is unseen during pre-training of the LLM. The results for Thai show variability. While dialogue history and contrastive learning yield minor improvements, integrating all context types with contrastive learning results in a decline. This pattern suggests that in languages characterized by tonal complexity and noisier biasing terms, the simultaneous alignment of multiple context sources can amplify recognition errors rather than mitigate them.

Overall, the results indicate that contextual information substantially improves multilingual ASR, though its effectiveness varies with the type of context and the target language. Contrastive learning strengthens the use of dialogue history but becomes less stable when heterogeneous contexts are introduced. The contrasting results show that effective context-aware ASR requires careful control of biasing terms and prompt design to avoid instability and preserve the benefits of contextual modeling.

5. Conclusion

This paper introduces a multilingual ASR framework designed to integrate contextual information into speech recognition without modifying the underlying speech encoder or language model. By combining a frozen speech encoder, a decoder-only LLM, and a lightweight projection module, the system supports various types of contextual input, such as dialogue history and biasing words, in a modular and efficient way.

We evaluate the proposed model on a real-world dataset, covering 11 different languages and multiple English accents. The results show that contextual information consistently improves recognition across all languages and accents, with an average error rate reduction of over 5% compared to the no-context baseline. Contrastive learning further enhances performance when applied to different context types, especially dialogue history, which achieves the best overall results. For instance, the combination of history and contrastive learning yields the lowest average error rate across settings, with substantial improvements in German, Korean, and Portuguese. Surprisingly, combining both dialogue history and biasing words with contrastive learning does not lead to the best performance and in some cases slightly underperforms the all context-only setting. This suggests that while contextual cues and alignment objectives are beneficial, their interaction in multi-context setups can introduce interference, highlighting the need for more adaptive integration strategies.

6. Ethics Statements

This work focuses on improving multilingual ASR through context-aware generation and alignment between speech and language modalities. All experiments were conducted using the publicly available MLC-SLM dataset containing real-world, consented multilingual speech data spanning diverse languages and dialects. We ensured that no personally identifiable information or sensitive content was used in training or evaluation. Additionally, care should be taken when deploying the system in high-stakes applications, where misrecognition or inappropriate use of contextual prompts could lead to biased or misleading outputs.

7. Limitations

This work investigates context-aware multilingual ASR using two specific types of context: dialogue history and biasing words. While these choices cover common conversational and domain-specific

cues, other potentially useful context signals, such as speaker identity, acoustic environment, or visual grounding, are not explored and may further improve transcription quality. Additionally, our evaluation is limited to the MLC-SLM dataset using EuroLLM. Generalization to under-represented languages, out-of-domain distributions, and low-resource or noisy conditions remains an open question. Future work could explore broader contextual types and extend evaluation to additional pretrained models and multilingual datasets.

8. Acknowledgments

This work was supported by the ELOQUENCE project (grant number 101070558) funded by the UKRI and the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the UKRI, European Union, or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. 2021. Context-aware transformer transducer for speech recognition. In *2021 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 503–510. IEEE.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

Jian Cheng. 2024. Context-aware speech recognition using prompts for language learners. *Interspeech 2024*, pages 4009–4013.

Lorenzo Concina, Jordi Luque, Alessio Brutti, Marco Matassoni, and Yuchen Zhang. 2025. The eloquence team submission for task 1 of mlc-slm challenge. *arXiv preprint arXiv:2507.19308*.

Ruchao Fan, Bo Ren, Yuxuan Hu, Rui Zhao, Shujie Liu, and Jinyu Li. 2025. Alignformer: Modality matching can achieve better zero-shot instruction-following speech-llm. *IEEE Journal of Selected Topics in Signal Processing*.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Xuandi Fu, Kanthashree Mysore Sathyendra, Ankur Gandhe, Jing Liu, Grant P Strimel, Ross McGowan, and Athanasios Mouchtaris. 2023. Robust acoustic and semantic contextual biasing in neural transducers for speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hayato Futami, Emiru Tsunoo, Yosuke Kashiwagi, Hiroaki Ogawa, Siddhant Arora, and Shinji Watanabe. 2024. Phoneme-aware encoding for prefix-tree-based contextual asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10641–10645. IEEE.

Xun Gong, Anqi Lv, Zhiming Wang, and Yanmin Qian. 2024. Contextual biasing speech recognition in speech-enhanced large language model. *Proc. Interspeech. ISCA*, pages 257–261.

Dongyue Guo, Zichen Zhang, Peng Fan, Jianwei Zhang, and Bo Yang. 2021. A context-aware language model to improve the speech recognition in air traffic control. *Aerospace*, 8(11):348.

Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. 2024. Integrating pre-trained speech and language models for end-to-end speech recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13289–13305.

Kaixun Huang, Ao Zhang, Zhanheng Yang, Pengcheng Guo, Bingshen Mu, Tianyi Xu, and Lei Xie. 2023. Contextualized end-to-end speech

recognition with contextual phrase prediction network. In *Proc. Interspeech 2023*, pages 4933–4937.

Rongqing Huang, Ossama Abdel-Hamid, Xinwei Li, and Gunnar Evermann. 2020. Class lm and word mapping for contextual biasing in end-to-end asr. *arXiv preprint arXiv:2007.05609*.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, et al. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. In *Proc. Interspeech 2021*, pages 1772–1776.

Zehan Li, Yan Yang, Xueqing Li, Jian Kang, Xiao-Lei Zhang, and Jie Li. 2025. Multilingual speech recognition using discrete tokens with a two-step training strategy. *arXiv preprint arXiv:2509.01900*.

Julian Linke, Jana Winkler, and Barbara Schuppler. 2025. Context is all you need? low-resource conversational asr profits from context, coming from the same or from the other speaker. In *Interspeech 2025*.

Loren Lugosch, Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan Collobert. 2022. Pseudo-labeling for massively multilingual speech recognition. *ICASSP*.

Nam Luu and Ondřej Bojar. 2025. End-to-end automatic speech recognition and speech translation: Integration of speech foundational models and llms. *arXiv preprint arXiv:2510.10329*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Bingshen Mu, Pengcheng Guo, Zhaokai Sun, Shuai Wang, Hexin Liu, Mingchen Shao, Lei Xie, Eng Siong Chng, Longshuai Xiao, Qiangze Feng, et al. 2025. Summary on the multilingual conversational speech language model challenge: Datasets, tasks, baselines, and methods. *arXiv preprint arXiv:2509.13785*.

Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. Deep context: end-to-end contextual speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, pages 418–425. IEEE.

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

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sébastien Bratières, et al. 2024. How to connect speech foundation models and large language models? what matters and what does not. *arXiv preprint arXiv:2409.17044*.

Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Ctc-assisted llm-based contextual asr. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 126–131. IEEE.

Maike Züfle and Jan Niehues. 2024. Contrastive learning for task-independent speechllm-pretraining. *arXiv preprint arXiv:2412.15712*.