CMC-SC: Cross-Modal Contextualized ASR Spelling Correction via BERT and WavLM using a Soft Fusion Framework

Mohammad Reza Peyghan and **Sajjad Amini*** and **Shahrokh Ghaemmaghami**
Electronics Research Institute and Department of Electrical Engineering
Sharif University of Technology
{m.peyghan, s_amini, ghaemmag}@sharif.edu

Abstract

Automatic Speech Recognition (ASR) systems remain error-prone in challenging acoustic conditions, leading to spelling mistakes that degrade downstream applications. Despite the surge in the number of studies on post-refinement methods, existing Spelling Correction (SC) approaches often rely solely on textual cues or phonetic features, limiting their ability to provide speech-aware corrections. In this work, we introduce a Cross-Modal Contextualized Spelling Correction framework (CMC-SC) that jointly incorporates contextualized acoustic and textual information. Unlike prior methods that use phonetics solely for candidate selection, our solution leverages contextualized speech tokens in the generation of corrections, improving accuracy and context awareness. CMC-SC features a detection module to identify errors, a cross-modal correction module to generate fixes using acoustic and textual tokens, and a soft fusion step to refine corrections while retaining context. The proposed method improves error rates compared to baselines and, with only 140M trainable parameters, offers an efficient solution for ASR spelling correction.

1 Introduction

Automatic Speech Recognition (ASR) systems have become increasingly important in recent years, enabling a wide range of applications, from virtual assistants to transcription services. The field has seen significant growth, driven by advancements in deep learning and natural language processing. However, despite these advances, ASR systems still face challenges, particularly in diverse acoustic environments and with speakers of different accents Errattahi et al. (2018). Retraining ASR models with domain-specific data can often mitigate these issues to some extent, but in many cases, the ASR model is not accessible for direct modification, functioning as a black box. In such scenarios, post-refinement techniques can be effectively employed to improve transcription quality.

Various ASR refinement techniques have been explored, especially since the advent of Transformers Vaswani et al. (2017). Broadly speaking, ASR refinement can be categorized into three main classes: fusion, re-scoring, and correction.

Fusion methods aim to improve ASR first-pass decoding by integrating external linguistic information at each decoding step. These techniques typically augment the ASR decoder's internal language model with external Language Models (LMs), whether a simple n-gram Kannan et al. (2018), a neural LM Kim et al. (2021), or a Large Language Model (LLM) Hori et al. (2025).

The Re-Scoring paradigm, by contrast, is a second-pass scheme that assumes the 1-best ASR hypothesis may not properly represent the information from the decoding step. This paradigm generates an N-best list of hypotheses and uses an external model (e.g., an n-gram or neural language model) to re-rank those candidates, selecting a linguistically superior candidate Shin et al. (2019); Gandhe and Rastrow (2020).

Correction approaches tackle the problem by revising a given ASR transcript to produce a new, improved sequence. Some correction techniques employ a second-pass decoding strategy, where a second decoder (or encoder-decoder) reconsiders acoustic features or the initial hypothesis. This decoding step can utilize an n-gram Bassil and Semaan (2012), a neural LM Zhang et al. (2019), or an LLM Udagawa et al. (2024), whether adopting both modalities Orihashi et al. (2021); Xia et al. (2017) or text-only correction Hrinchuk et al. (2020); Jia et al. (2025). In recent research, researchers have used Retrieval Augmented Gen-

* Corresponding author: s_amini@sharif.edu
Code available at:
https://github.com/mohammadr8za/CMC-SC.git

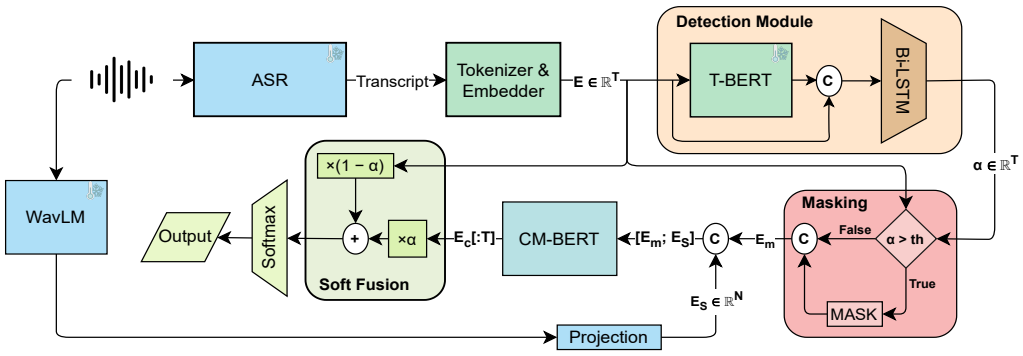

Figure 1: The Overall Diagram of the Proposed CMC-SC.

eration (RAG) with an external corpus for transcript correction Robatian et al. (2025); Gong et al. (2025).

Another notable approach to ASR error correction is the use of encoder transformers, primarily BERT Devlin et al. (2019). These models leverage their contextual understanding to replace erroneous tokens. However, using pre-trained BERT alone is suboptimal for ASR correction Zhang et al. (2020) due to (1) reliance on textual cues, which risks incorrect substitutions, and (2) a domain mismatch between its clean pre-training data and noisy ASR outputs. To address this, FASPell Hong et al. (2019) employs a Confidence-Similarity Decoder (CSD) to filter BERT's candidates by phonetic and orthographic similarity. Similarly, SpellGCN Cheng et al. (2020) enhances BERT with a Graph Convolutional Network (GCN) to model phonological and symbolic relations. Other methods incorporate detection modules. For instance, a method Zhang et al. (2020) detects and softly masks probable errors based on confidence scores, feeding them into a correction model and summing outputs with original embeddings. Another approach Zhang et al. (2021) fuses token and phonetic embeddings post-detection for phonetic-aware correction. Additionally, a dynamic error scaling method Fan et al. (2023) integrates words and pinyin for semantically and phonetically aware character-level correction.

However, encoder-based methods addressing these challenges often rely on phonetic information derived from text, which can be misleading. Our method addresses this issue by:

1. Extracting contextualized acoustic information directly from speech using WavLM Chen et al. (2022), unlike Fan et al. (2023) and Zhang et al. (2021), which rely on transcription-based information.

2. Joint processing of contextualized acoustic and textual tokens through a Cross-Modal BERT (CM-BERT) unlike Hong et al. (2019); Cheng et al. (2020), which rely on phonetic information in a secondary branch.

3. Using a soft fusion technique to combine CM-BERT outputs with original token embeddings, preserving transcription information, unlike the direct summing approach used in Zhang et al. (2021).

Finally, our approach improves upon existing baselines by a large margin, demonstrating its effectiveness in improving ASR quality.

2 Method

This section presents the methodology for enhancing ASR transcriptions using a cross-modal framework that integrates textual and acoustic data. The approach comprises two main components: a detection module to identify erroneous tokens and a cross-modal correction module to rectify these errors using a soft-fusion framework. The structure of the proposed CMC-SC is illustrated in Figure 1, and subsequent subsections detail each component.

2.1 Data Pre-Processing

We perform the following data preprocessing steps to enable end-to-end (E2E) training of our proposed model:

1. We run paired speech-text examples through a black-box ASR to obtain its transcriptions.

2. For each utterance, we align the ASR transcription with the corresponding ground-truth using Levenshtein alignment (edit distance).

From this, we create a per-token binary sequence where '1' indicates an erroneous token and '0' indicates a correctly recognized token.

Consequently, our dataset comprises samples in the form of (utterance, erroneous text, labels, target text). The (erroneous text, labels) pairs are used to train the Detection module, while the (utterance, erroneous text, target text) pairs are used to train the Correction module.

2.2 Detection and Masking

The detection module, the first stage of our spelling correction pipeline, aims to identify erroneous tokens in ASR transcriptions to enable targeted corrections and prevent over-correction, which is a common issue in E2E methods Imai et al. (2025). It integrates a frozen and pre-trained BERT model (denoted T-BERT) to extract contextual token embeddings, combines these embeddings with initial token embeddings via a residual connection, and feeds them to a two-layer BiLSTM classification head. The residual connection is crucial, as it allows the model to consider content other than context, addressing potential misclassifications from incorrect tokens affecting the frozen BERT's embeddings. Then, a linear layer outputs logits for each token, indicating whether it is erroneous. This module is trained using Binary Cross-Entropy (BCE) loss with logits and per-token binary labels:

$$\mathrm{BCE}_D = \frac{1}{T}\sum_{i=1}^{T} \log\bigl(1 + e^{-(2y_i-1)z_i}\bigr).$$

where z_i is the logit for token i, and $y_i \in \{0, 1\}$ indicates if the token is erroneous.

At inference time, we compute the sigmoid of each logit and compare it to a predefined threshold; tokens with likelihood above this threshold are deemed erroneous and replaced with the `[MASK]` token. This masking strategy ensures that CM-BERT's context is derived from the most probable correct tokens, preventing incorrect tokens from negatively affecting the contextual representation.

2.3 Soft Fusion and Cross-Modal Correction

The cross-modal correction module refines ASR transcriptions by integrating textual and acoustic data to produce accurate, speech-aware, and contextually appropriate corrections. Using both modalities, it improves transcription quality using a cross-modal and joint attention approach.

The correction module receives a sequence of token embeddings from the detection phase, where tokens identified as incorrect are replaced with the `[MASK]` token, denoted as $\mathbf{E}_m$. It also extracts contextualized speech features from raw audio using a pre-trained WavLM network. These speech features are then projected to match the dimensionality of the CM-BERT, resulting in $\mathbf{E}_S$.

The masked text embeddings $\mathbf{E}_m$ and the projected speech embeddings $\mathbf{E}_S$ are concatenated to form the input $\mathbf{E}_{\text{in}} = [\mathbf{E}_m; \mathbf{E}_S]$. This concatenated input is then processed by the CM-BERT, a transformer-based model that outputs contextualized, speech-aware representations by enabling cross-modal interactions between text and speech through its attention mechanisms.

To prevent over-correction and preserve correct tokens, the Soft-Fusion (SF) strategy blends each token's original embedding $\mathbf{E}^{(i)}$ with its corresponding cross-modal contextual embedding $\mathbf{E}_c^{(i)}$, based on a confidence score α_i from the detection phase that indicates the likelihood that token i is incorrect. Specifically, under the SF strategy, the output embedding for each token i is computed as:

$$\mathbf{E}_o^{(i)} = (1 - \alpha_i) \cdot \mathbf{E}^{(i)} + \alpha_i \cdot \mathbf{E}_c^{(i)}$$

As a result, tokens with a low α_i (indicating they are likely correct) retain more of their original embedding, while tokens with a high α_i (indicating they are likely incorrect) incorporate more of the speech-informed representation. This adaptive interpolation ensures precise corrections where needed while preserving accurate text.

Finally, the softly-fused embeddings are classified into tokens using a softmax layer, guided by the Cross Entropy (CE) loss with logits and token IDs. The CE loss is given by:

$$\mathrm{CE}_C = -\frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{\exp(z_{t,y_t})}{\sum_{k=1}^{V}\exp(z_{t,k})}\right)$$

where T is the sequence length, V is the vocabulary size, $z_{t,k}$ is the logit for token k at position t, and y_t is the true token ID.

3 Experiments and Results

In this section, we detail the experiments conducted to assess the proposed model and present the results in comparison to several baseline models, which were re-implemented to ensure a fair evaluation. Additionally, we assess the performance of each

	Model	Parameters		Detection			Correction			Error Rate	
		Total	Trainable	P	R	F1	P	R	F1	Word	Character
Comparative Study	Whisper-Tiny (Baseline) Radford et al. (2023)	39M	-	-	-	-	-	-	-	24.5	17.2
	Whisper-Small Radford et al. (2023)	244M	-	-	-	-	-	-	-	13.7	6.1
	Whisper-Medium Radford et al. (2023)	769M	-	-	-	-	-	-	-	11.7	**4.2**
	PT-BERT+BiLSTM (Multi-Task Training)	140M	30M	85.96	85.90	85.88	81.74	83.94	81.82	18.2	17.5
	FT-BERT+BiLSTM (Multi-Task Training)	140M	140M	85.84	85.95	85.79	82.08	85.77	82.18	17.9	17.1
	Soft-Masked BERT Zhang et al. (2020)	250M	250M	86.14	86.23	86.12	87.85	87.23	87.18	13.2	9.8
	CMC (Ours)	300M	140M	**87.32**	**87.52**	**87.30**	**91.31**	**91.35**	**91.27**	**9.2**	5.6
Ablation Study	CMC – WavLM	210M	140M	87.18	87.24	87.15	88.91	88.74	88.64	12.1	9.6
	CMC – SF	300M	140M	87.17	87.18	87.15	89.03	89.12	89.01	10.7	7.5

Table 1: Comparative and Ablation Studies (all refinement methods are applied to Whisper-Tiny)

module within the model through an ablation study, systematically removing each module to evaluate its impact on the overall performance.

To evaluate our model, we introduce baseline models. We use three ASR models (Whisper-Tiny, Whisper-Small, Whisper-Medium) to assess the importance of post-refinement and Cross-Modal attention against adopting larger ASR systems. We also trained two spelling correction baselines, Pre-Trained (PT) and Fine-Tuned (FT) BERT, following Zhang et al. (2020); Cheng et al. (2020); Fan et al. (2023), to highlight our model's contribution. Plus, we re-implemented Soft-Masked BERT Zhang et al. (2020) as another benchmark.

We perform an ablation study to quantify each module's contribution to the CMC-SC model. First, we remove the speech tokens (i.e., contextualized acoustic information) and retrain under identical conditions, noting that CM-BERT is originally pre-trained on text, so its performance may still reflect textual bias rather than a true absence of cross-modal data. This ablation also underscores the significance of the residual connection in the detection module, which is the primary distinction of this module in the ablation and compared to the PT-BERT. Next, we remove the Soft-Fusion module, which retains information from the original transcription, and train it again. Table 1 presents these results, demonstrating that each module positively impacts the overall performance of CMC-SC.

All experiments ran on an NVIDIA RTX 3090 GPU for 30 epochs using the AdamW optimizer. The best model uses a batch size of 32, a learning rate of 1×10^{-5} with a linear scheduler, both T-BERT and CM-BERT have a maximum context length of 128 tokens, and the threshold in the masking module is set empirically to 0.5. To align speech tokens with BERT embeddings, we project them into 50 tokens of dimension 768. We have employed the Mozilla Common Voice dataset Ardila et al. (2019) (original train/dev/test splits) and report results on its test set.

Transcriptions	1) the cut start on the fence 2) she begged home smiling all the way knowing that she had won
Detection Predictions	1) [1, 1, 1, 0, 0, 0] 2) [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Detection Labels	1) [0, 1, 1, 0, 0, 0] 2) [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Refined	1) the cat sat on the fence 2) she biked home smiling all the way knowing that she had won
Ground Truths	1) the cat sat on the fence 2) she biked home smiling all the way knowing that she had won

Table 2: Examples of CMC-SC on the Common Voice test set.

Finally, as shown in Table 1, our proposed method improves the baselines by a large margin, demonstrating substantial potential to improve the spelling correction task. Notably, our model has only 140M trainable parameters and outperforms the pre-trained Whisper-medium with 769M parameters, making it a lightweight yet effective solution. The examples of CMC-SC are provided in Table 2.

4 Conclusion

In this paper, we have introduced Cross-Modal Contextualized Spelling Correction (CMC-SC), a novel framework designed to enhance ASR transcription accuracy by correcting spelling errors. CMC-SC integrates a detection module using a frozen BERT model and BiLSTM to identify errors by capturing contextual and sequential patterns, and a correction module that blends text embeddings with acoustic features from a pretrained

WavLM. This approach ensures precise, context-aware corrections while preserving accurate tokens via a soft fusion framework. Experiments show CMC-SC reduces error rates with only 140 million trainable parameters, balancing performance and computational efficiency. Future work includes supporting additional languages and integrating advanced pretrained cross-modal networks for deeper linguistic and acoustic insights.

Limitations

Despite the resulting advancements, ASR models remain error-prone in challenging environments. In clean settings, errors are primarily substitutions or spelling mistakes, for which spelling correction methods are computationally efficient. However, the proposed method may be less effective for errors involving insertions and deletions. Additionally, trained on general data, the model may require re-training for domain-specific applications, such as medical terminology.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Youssef Bassil and Paul Semaan. 2012. Asr context-sensitive error correction based on microsoft n-gram dataset. *arXiv preprint arXiv:1203.5262*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37.

Jiaxin Fan, Yong Zhang, Hanzhang Li, Jianzong Wang, Zhitao Li, Sheng Ouyang, Ning Cheng, and Jing Xiao. 2023. Boosting chinese asr error correction with dynamic error scaling mechanism. In *Proc. Interspeech 2023*, pages 2173–2177.

Ankur Gandhe and Ariya Rastrow. 2020. Audio-attention discriminative language model for asr rescoring. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7944–7948. IEEE.

Xun Gong, Anqi Lv, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. Br-asr: Efficient and scalable bias retrieval framework for contextual biasing asr in speech llm. *arXiv preprint arXiv:2505.19179*.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. Faspell: A fast, adaptable, simple, powerful chinese spell checker based on dae-decoder paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169.

Takaaki Hori, Martin Kocour, Adnan Haider, Erik McDermott, and Xiaodan Zhuang. 2025. Delayed fusion: Integrating large language models into first-pass decoding in end-to-end speech recognition. *arXiv preprint arXiv:2501.09258*.

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. 2020. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 7074–7078. IEEE.

Saki Imai, Tahiya Chowdhury, and Amanda Stent. 2025. Evaluating open-source asr systems: Performance across diverse audio conditions and error correction methods. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5027–5039.

Linzhao Jia, Han Sun, Yuang Wei, Changyong Qi, and Xiaozhe Yang. 2025. Epic: Error pattern informed correction for classroom asr with limited labeled data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE.

Suyoun Kim, Yuan Shangguan, Jay Mahadeokar, Antoine Bruguier, Christian Fuegen, Michael L Seltzer, and Duc Le. 2021. Improved neural language model fusion for streaming recurrent neural network transducer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7333–7337. IEEE.

Shota Orihashi, Ryo Masumura, Mana Ihori, Takafumi Moriya, Akihiko Takashima, Naoki Makishima, Takanori Ashihara, and Tomohiro Tanaka. 2021. Cross-modal transformer-based neural correction models for automatic speech recognition. *Interspeech 2021*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Amin Robatian, Mohammad Hajipour, Mohammad Reza Peyghan, Fatemeh Rajabi, Sajjad Amini, Shahrokh Ghaemmaghami, and Iman Gholampour. 2025. Gec-rag: Improving generative error correction via retrieval-augmented generation for automatic speech recognition systems. *arXiv preprint arXiv:2501.10734*.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using bert for speech recognition. In *Asian Conference on Machine Learning*, pages 1081–1093. PMLR.

Takuma Udagawa, Masayuki Suzuki, Masayasu Muraoka, and Gakuto Kurata. 2024. Robust asr error correction with conservative data filtering. *arXiv preprint arXiv:2407.13300*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. *Advances in neural information processing systems*, 30.

Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021. Correcting chinese spelling errors with phonetic pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2250–2261.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890.

Shiliang Zhang, Ming Lei, and Zhijie Yan. 2019. Investigation of transformer based spelling correction model for ctc-based end-to-end mandarin speech recognition. In *Interspeech*, pages 2180–2184.