Abjad AI at NADI 2025: CATT-Whisper: Multimodal Diacritic Restoration Using Text and Speech Representations

Ahmad Ghannam, Naif Alharthi, Faris Alasmary, Kholood Al Tabash, Shouq Sadah, and Lahouari Ghouti
Abjad AI. King Khaled Road. Riyadh 11000, Saudi Arabia.
{aghannam,nalharthi,falasmary,kaltabash,ssadah,lghouti}@abjad.com.sa

Abstract

In this work, we tackle the Diacritic Restoration (DR) task for Arabic dialectal sentences using a multimodal approach that combines both textual and speech information. We propose a model that represents the text modality using an encoder extracted from our own pre-trained model named CATT. The speech component is handled by the encoder module of the OpenAI Whisper base model. Our solution is designed following two integration strategies. The former consists of fusing the speech tokens with the input at an early stage, where the 1500 frames of the audio segment are averaged over 10 consecutive frames, resulting in 150 speech tokens. To ensure embedding compatibility, these averaged tokens are processed through a linear projection layer prior to merging them with the text tokens. Contextual encoding is guaranteed by the CATT encoder module. The latter strategy relies on cross-attention, where text and speech embeddings are fused. The cross-attention output is then fed to the CATT classification head for token-level diacritic prediction. To further improve model robustness, we randomly deactivate the speech input during training, allowing the model to perform well with or without speech. Our experiments show that the proposed approach achieves a word error rate (WER) of 0.25 and a character error rate (CER) of 0.9 on the development set. On the test set, our model achieved WER and CER scores of 0.55 and 0.13, respectively.

1 Introduction

Diacritics are essential for accurate interpretation, pronunciation, and meaning in Arabic. However, in most informal writing such as social media, messaging, or transcribed speech they are omitted. While native speakers often infer the intended forms from context, the absence of diacritics introduces significant ambiguity, particularly in dialects where phonetic and morphological variation is high and orthographic conventions are inconsistent. This not only challenges human readers but also degrades the performance of downstream NLP tasks such as speech synthesis, machine translation, and information retrieval. The NADI 2025 shared task overview (Talafha et al., 2025) highlights that DR remains particularly difficult for dialectal Arabic due to limited annotated data, regional variability, and inconsistent spelling practices. Traditional DR approaches rely solely on text, ranging from rule-based systems and n-gram models to transformer-based language models such as BERT (Devlin et al., 2019). These methods often fail when orthographic cues alone are insufficient, an issue exacerbated in dialectal and code-switched text. In contrast, speech carries prosodic and phonetic signals that can directly disambiguate diacritic placement, offering a valuable complement to text.

In this work, we propose CATT-Whisper, a multimodal DR system that integrates a CATT (Alasmary et al., 2024) text encoder with the Whisper (Radford et al., 2023) speech encoder. We evaluated two fusion strategies: **(i) Early fusion**: projected speech embeddings are merged with text embeddings before passing them to CATT encoder as inputs. **(ii) Cross-attention fusion**: the output of the CATT encoder is fused with the speech embeddings from Whisper using cross attention layer, followed by the classification layer.

Our contributions are: (i) A multimodal DR system for Arabic dialects combining large-scale pre-trained text and speech encoders. (ii) Comparative analysis of early fusion vs. Cross-attention fusion. (iii) A modality-robust training scheme for variable speech availability. Our full codebase, including pre-trained models and training scripts, is publicly available [1], ensuring reproducibility and facilitating further research in multimodal DR.

[1] https://github.com/abjadai/catt-whisper

Proceedings of The Third Arabic Natural Language Processing Conference, pages 757–761

2 Background

2.1 Task Setup

The DR shared subtask at NADI 2025 focuses on restoring missing diacritics in Arabic text, with the option to also use speech for better performance. Unlike most previous work that only targets MSA, this task also covers Classical Arabic, dialects, and code-switched text, which are more challenging. Some examples are provided in Table 1.

Example Input 1	عندكو شوربة ايه النهرده
CATT	عِنْدَكُو شُورْبَةُ ايه النَّهْرَدَهْ
CATT-Whisper	عَندُكُو شوربِة اِيه النِهَرْدَه
Reference	عَندُكُو شوربِة اِيه النِهَرْدَه
Example Input 2	عايز شوية وأت لتجهيز الاكل
CATT	عَايَزَ شُوِيَّةً وَأْتْ لِتَجْهِيزِ الاكْلِ
CATT-Whisper	عَايِز شوَيَّة وَأت لِتَجهِيز الاَكل
Reference	عَايِز شوَيَّة وَأت لِتَجهِيز الاَكل

Table 1: Examples from the NADI 2025 Subtask 3 dataset (dev/test). CATT (text-only) and CATT-Whisper (speech-enhanced) outputs compared with references, showing how speech features resolve phonological ambiguities.

2.2 Dataset

Our experiments were conducted using the NADI 2025 DR dataset, provided as part of the shared task, which is publicly available on Hugging Face [2]. The dataset covers a mix of dialectal, multi-dialectal, and Classical Arabic varieties, with some segments exhibiting code-switching between Arabic and other languages. The dataset is a combined collection derived from several resources, namely MDASPC (Almeman et al., 2013), TunSwitch (Abdallah et al., 2023), ArzEn (Hamed et al., 2020), Mixat (Al Ali and Aldarmaki, 2024), ClArTTS (Kulkarni et al., 2023), and ArVoice (Toyin et al., 2025). While the CATT and Whisper models we use in our system were already pretrained on their respective large-scale corpora, the NADI 2025 DR dataset used exclusively for fine-tuning the combined architecture for this DR task. The provided training data consists of multiple sub-datasets, summarized in Table 2.

[2] https://huggingface.co/datasets/MBZUAI/NADI-2025-Sub-task-3-all

Dataset	Type	Dia.	Train
MDASPC	Multi-dialectal	True	60,677
TunSwitch	Dialectal, CS	True	5,212
ArzEn	Dialectal, CS	False	3,344
Mixat	Dialectal, CS	False	3,721
ClArTTS	CA	True	9,500
ArVoice	MSA	True	2,507

Table 2: Statistics of the NADI 2025 Subtask 3 datasets. CA = Classical Arabic, CS = Code-Switched Arabic, Dia. = diacritic. The table reports the number of sentences in each split.

2.3 Related Work

Research on Arabic DR has evolved from rule-based methods to neural and multimodal approaches (Elgamal et al., 2024). Early systems relied on lexicons and morphological analyzers, later extended with n-gram models (Habash and Rambow, 2007; Elshafei et al., 2006), but they struggled with dialectal variation, noisy text, and borrowed vocabulary. Neural models, from RNNs and LSTMs (Zitouni et al., 2006; Belinkov and Glass, 2015) to transformers (Nazih and Hifny, 2022) with pre-trained language models such as AraBERT (Antoun et al., 2020), CAMeLBERT (Inoue et al., 2021), and CATT (Alasmary et al., 2024), improved accuracy but still failed to resolve phonetic ambiguities in dialects. While (Elgamal et al., 2024) highlighted the usefulness of "diacritics-in-the-wild" signals, text-only models remain insufficient for ambiguous cases.

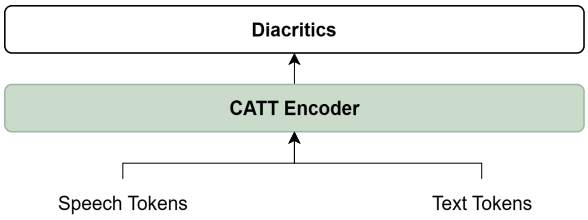
Multimodal approaches increasingly exploit ASR outputs as phonetic cues. Early work (Aldarmaki and Ghannam, 2023) relies solely on ASR, which can produce both transcripts and diacritic predictions, but errors in transcription often propagate to diacritization. More recent methods (Shatnawi et al., 2024a) integrate ASR-derived-diacritized transcripts with undiacritized text via cross-attention, enhancing performance while still being sensitive to ASR noise.

Our approach differs by (i) deeply integrating text and speech through early and cross-attention fusion, (ii) focusing explicitly on dialectal DR with robust pre-trained encoders: CATT and Whisper.

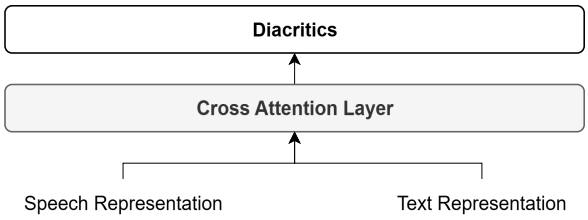
3 System Overview

3.1 Architecture Components

The architecture consists of a **Text Encoder**, implemented with a pre-trained CATT model for DR, and



(a) Early Fusion



(b) Cross Attention Fusion

Figure 1: Proposed CATT-Whisper Architectures for Multimodal. (a) Early Fusion Configuration. (b) Cross-Attention Fusion Configuration.

a **Speech Encoder**, implemented with the Encoder part of Whisper-Base model. A **Linear Projection Layer** follows the speech encoder to match the dimensionality of the text encoder. The proposed architectures are summarized in Figure 1.

3.2 Fusion Strategies

3.2.1 Early Fusion

Speech features are downsampled from 1,500 frames to 150 tokens by averaging 10 frames with and projecting them to match the text embedding dimension. These speech tokens are then concatenated with text tokens and fed into the CATT encoder, following a strategy similar to (Wu et al., 2023). This early fusion approach can be seen as a form of "soft prompting," where text tokens are augmented with speech embeddings via speech-placeholder tokens, enabling the model to leverage acoustic features while preserving the core CATT architecture. Details of this fusion strategy is shown in Figure 2.

3.2.2 Cross-Attention Fusion

Text and speech embeddings are encoded separately, then fused via a cross-attention layer before being passed to the classification layer, similar to the multi-modal setup of (Shatnawi et al., 2024b).

3.2.3 Fusion Strategy Choice

In our experiments, both Early Fusion and Cross-Attention Fusion yielded comparable results. However, as Cross-Attention is computationally more demanding, we focused on Early Fusion, and all results reported in this paper correspond to this configuration.

3.3 Speech Augmentation

Time-frequency warping (Park et al., 2019) is applied during training to improve generalization.

4 Experimental Setup

For training, we used the NADI 2025 DR train and development sets, while evaluation was performed on the official test set. Model performance was measured using Word Error Rate (WER) and Character Error Rate (CER), which are the standard metrics. Our preprocessing step included tokenization, speech feature extraction, and spectrogram augmentation through time-frequency warping.

Training was carried out with a batch size of 32, a learning rate of 1×10^{-5}, a dropout rate of 0.1, and the AdamW optimizer. During training, the speech encoder was frozen for the first 5 epochs allowing the projection layer to adapt, then unfrozen and jointly trained with the rest of the model for more 5 epochs. This two-phase procedure was applied in all experiments for both fusion models.

5 Results

5.1 Development Set Performance

Table 3 shows the performance of our proposed model compared to other works on the development set. Our model achieves substantially lower word error and character error rates (WER and CER).

Participant	WER	CER
gahmed92 (Ours)	**0.25**	**0.09**
omarnj	0.46	0.22
Baseline	0.46	0.22

Table 3: Results on the NADI 2025 Subtask 3 official development set, reported in WER and CER

5.2 Test Set Performance

Table 4 presents the results on the official test set. Our models outperforms all models in both metrics.

5.3 Performance on Challenging Test Cases

We further analyzed the model on a set of challenging test cases recorded by our team, where the

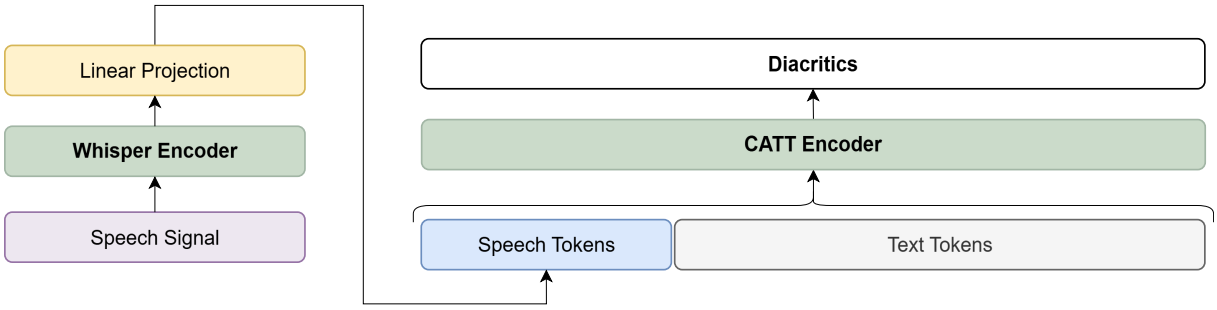


Figure 2: Early Fusion architecture of the proposed CATT-Whisper model. Speech features are downsampled and projected to match text embeddings before being concatenated with text tokens and processed by the CATT encoder.

Participant	WER	CER
gahmed92 (Ours)	**0.55**	**0.13**
mohamed_elrefai	0.64	0.15
Baseline	0.65	0.16

Table 4: Results on the NADI 2025 Subtask 3 official test set, reported in WER and CER

same word is pronounced differently within the same sentence. The results, summarized in Table 5, show that while our model achieves lower WER and CER than the others, these cases remain difficult and are not fully solved. This highlights both the robustness of our approach and the need for further improvements to handle complex, real-world pronunciation variability.

Example Input 1	ضرب ضرب ضرب
CATT-Whisper	ضَرِب ضَرِب ضَرِب
Reference	ضَرَبَ ضُرِبَ ضَرْبٌ
Example Input 2	ذهب ذهب
CATT-Whisper	ذَهِب ذَهِب
Reference	ذَهَبٌ ذَهَبْ

Table 5: Model performance on challenging test cases with variable word pronunciations.

6 Conclusion

We present CATT-Whisper, a multimodal system for Arabic DR that combines pre-trained text and speech encoders via early fusion and cross-attention. Both strategies achieve competitive results. While speech input boosts diacritic accuracy, some ambiguous sequences remain challenging, suggesting the need for stronger phoneme-level encoders (e.g., CTC-based models such as Conformer-CTC (Gulati et al., 2020), Squeezeformer (Kim et al., 2022)). Future work will explore alternative acoustic models and larger-scale training.

Acknowledgments

The authors would like to thank the management of AbjadAI Company for their generous support.

References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2023. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. *Preprint*, arXiv:2309.11327.

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.

Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. 2024. CATT: Character-based Arabic tashkeel transformer. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 250–257, Bangkok, Thailand. Association for Computational Linguistics.

Hanan Aldarmaki and Ahmad Ghannam. 2023. Diacritic recognition performance in arabic asr. In *Interspeech 2023*, pages 361–365.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Yonatan Belinkov and James Glass. 2015. Arabic diacritization with recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2281–2285, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Salman Elgamal, Ossama Obeid, Mhd Kabbani, Go Inoue, and Nizar Habash. 2024. Arabic diacritics in the wild: Exploiting opportunities for improved diacritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14815–14829, Bangkok, Thailand. Association for Computational Linguistics.

Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Alghamdi. 2006. Statistical methods for automatic diacritization of arabic text. *The Saudi 18th National Computer Conference. Riyadh*, 18:301–306.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Nizar Habash and Owen Rambow. 2007. Arabic diacritization through full morphological tagging. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 53–56, Rochester, New York. Association for Computational Linguistics.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An efficient transformer for automatic speech recognition. In *Advances in Neural Information Processing Systems*, volume 35, pages 9361–9373. Curran Associates, Inc.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'em Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus. In *2023 INTERSPEECH*, pages 5511–5515.

Waleed Nazih and Yasser Hifny. 2022. Arabic syntactic diacritics restoration using bert models. *Computational Intelligence and Neuroscience*, 2022.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sara Shatnawi, Sawsan Alqahtani, and Hanan Aldarmaki. 2024a. Automatic restoration of diacritics for speech data sets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4166–4176, Mexico City, Mexico. Association for Computational Linguistics.

Sara Shatnawi, Sawsan Alqahtani, Shady Shehata, and Hanan Aldarmaki. 2024b. Data augmentation for speech-based diacritic restoration. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 160–169, Bangkok, Thailand. Association for Computational Linguistics.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. Nadi 2025: The first multidialectal arabic speech processing shared task. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Hawau Toyin, Rufael Marew, Humaid Alblooshi, Samar M. Magdy, and Hanan Aldarmaki. 2025. ArVoice: A Multi-Speaker Dataset for Arabic Speech Synthesis. In *Interspeech 2025*, pages 4808–4812.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Imed Zitouni, Jeffrey S. Sorensen, and Ruhi Sarikaya. 2006. Maximum entropy based restoration of Arabic diacritics. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 577–584, Sydney, Australia. Association for Computational Linguistics.