TantaArabNLP at KSAA-2026 Task 2: Adapting CATT-Whisper for Arabic Speech Dictation with Automatic Diacritization

Nada Esmaeil, Reda M. Elbasiony, Mohamed T. Faheem
Faculty of Engineering, Tanta University
Tanta, Egypt
{nada_adel, reda, mohamed_ahmed1}@f-eng.tanta.edu.eg

Abstract

We present our submission to the KSAA-2026 Shared Task (Subtask 2): Automatic Diacritization of Speech Dictation. Building upon the CATT-Whisper multimodal architecture, which fuses representations from a pre-trained CATT text encoder and the Whisper speech encoder, we fine-tune the model end-to-end on the official shared task training data. To further enhance performance on speech-dictated Arabic text, we apply careful post-processing to the model outputs. Our best submission achieves a Diacritic Error Rate (DER) of 7.04, a Word Error Rate (WER) of 24.39, and a Sentence Error Rate (SER) of 71.65 on the hidden test set, securing 2nd place in the competition. These results demonstrate the effectiveness of adapting a strong multimodal baseline to the speech-aware diacritization setting and highlight the value of task-specific fine-tuning and output refinement for bridging the gap between spoken transcripts and fully diacritized Arabic text.

Keywords: Arabic diacritization, Speech-aware diacritization, Multimodal diacritic restoration, CATT-Whisper

1. Introduction

The Arabic language is characterized by a rich diacritic system that includes short vowels (fatḥa, ḍamma, kasra), sukūn, shaddah, and tanwīn marks. These diacritics are essential for correct pronunciation, meaning, and grammatical structure. However, diacritics are routinely omitted in most written Arabic texts, including speech transcripts. This creates significant ambiguity for computational systems and downstream applications such as ASR, TTS, and machine translation.

Traditional text-only diacritization models perform well on formal MSA but struggle with speech transcripts, which are less formal and may include repetitions or colloquial elements. KSAA-2026 Subtask 2 introduces a *speech-aware diacritization* benchmark, requiring systems to combine raw audio and undiacritized transcripts to produce fully diacritized text.

In this work, we describe our submission to KSAA-2026 Subtask 2, building on the multimodal CATT-Whisper architecture (Ghannam et al., 2025). Our main contributions are:

- Fine-tuning CATT-Whisper Early Fusion on the official dataset using selective layer tuning with differential learning rates.
- A deterministic post-processing module to restore punctuation, numbers, and non-Arabic tokens.
- Achieving a DER of 7.04, WER of 24.39, and SER of 71.65 on the hidden test set, securing second place.

2. Related Work

2.1. Text-based Diacritic Restoration

Early Arabic diacritic restoration systems relied heavily on rule-based and morphological analyzers. Prominent examples include MADAMIRA (Pasha et al., 2014) and Camelira (Obeid et al., 2022), which perform diacritization as part of full morphological disambiguation. Subsequent research moved toward neural sequence labeling approaches using RNNs, BiLSTMs with CRFs (Althubaity et al., 2020), and transformer-based models. Among recent models, the Character-based Arabic Tashkeel Transformer (CATT) (Alasmary et al., 2024) has established itself as a strong character-level baseline for Arabic diacritization.

Despite notable progress on well-formed Modern Standard Arabic (MSA) text, text-only models often suffer from significant domain mismatch when applied to speech transcripts, which tend to be more informal, repetitive, or dialect-influenced (Aldarmaki and Ghannam, 2023).

2.2. Speech-aware and Multimodal Diacritic Restoration

The use of acoustic information to support diacritic restoration remains relatively underexplored. One of the earliest attempts dates back to Vergyri and Kirchhoff (Vergyri and Kirchhoff, 2004), who combined acoustic features with morphological analysis for dialectal Arabic. More recently, Aldarmaki and Ghannam (Aldarmaki and Ghannam, 2023) showed that text-based diacritizers generalize poorly to ASR outputs and that models trained

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 229–233
May 11, 2026. ©ELRA Language Resources Association (ELRA), 2026

on gold diacritized speech data achieve superior performance.

With the advancement of large pretrained models, multimodal approaches have begun to emerge. Shatnawi et al. (Shatnawi et al., 2024) explored cascaded pipelines in which a fine-tuned Whisper model first generates diacritized transcripts that are then refined by a text-based restoration model.

The most relevant work to our submission is the CATT-Whisper architecture (Ghannam et al., 2025), which integrates a CATT text encoder with the Whisper speech encoder using both early fusion and cross-attention mechanisms. Originally developed for dialectal Arabic diacritic restoration in the NADI 2025 shared task, CATT-Whisper demonstrated the effectiveness of combining textual and acoustic signals in a multimodal framework.

3. Background

3.1. Task Setup

The KSAA-2026 Shared Task (Subtask 2), titled *Automatic Diacritization of Speech Dictation*, requires participants to develop multimodal systems that take as input a raw Arabic speech audio file together with its corresponding undiacritized transcript and generate a fully diacritized version of the transcript at the character level. The model must predict all Arabic diacritic marks, including fatḥa, ḍamma, kasra, sukūn, shaddah, and tanwīn.

Systems are evaluated using three metrics: Diacritic Error Rate (DER), Word Error Rate (WER), and Sentence Error Rate (SER), with WER serving as the primary metric. Results are reported with and without case endings (i'rāb). Case endings constitute the most challenging aspect of Arabic diacritization as they heavily depend on syntactic context.

3.2. Dataset

The dataset was collected through the VoiceWall crowdsourcing platform developed by the King Salman Global Academy for Arabic Language (KSAA). It consists of approximately five hours of high-quality Arabic speech recordings from both male and female speakers, covering Modern Standard Arabic (MSA) as well as various Arabic dialects. All utterances are short (maximum 9 seconds) to ensure accurate speech–text alignment.

The official dataset splits are summarized in Table 1.

Split	# Files	Duration	Gender
Training	2,327	≈4.5 h	1,423M / 904F
Development	260	≈0.5 h	161M / 99F
Test	260	≈0.5 h	154M / 104F

Table 1: Statistics of the KSAA-2026 Subtask 2 dataset.

4. System Overview

We built our submission for KSAA-2026 Subtask 2 upon the CATT-Whisper multimodal architecture (Ghannam et al., 2025). In addition, we conducted experiments with the pure CATT model using encoder-decoder configuration. This section describes the architectural choices and the experiments we performed.

4.1. Baselines

The organizers provided three baseline systems as reference implementations:

- **Text-only**: A transformer model using only undiacritized text.
- **Speech + text**: A transformer model combining ASR outputs with undiacritized text, without raw acoustic features.
- **Fine-tuned**: A transformer model fine-tuned on the official KSAA-2026 training data.

4.2. Architectural Choices

4.2.1. CATT Text-only Model

We used the Encoder-Decoder (ED) variant of the pretrained CATT model (Alasmary et al., 2024). In this setup, the undiacritized Arabic text is fed as input and the model generates the full sequence of diacritics in an auto-regressive manner. This architecture conditions each predicted diacritic on both the input text and the previously generated diacritics, which is particularly helpful for modeling long-range dependencies.

4.2.2. CATT-Whisper Multimodal Model

Our best-performing system is based on the CATT-Whisper architecture (Ghannam et al., 2025). This model combines:

- A pretrained CATT model as the **text encoder**.
- The encoder part of the Whisper-Base model as the **speech encoder**.
- A linear projection layer to align the speech embeddings with the text embedding dimension.

5. Experimental Setup

We conducted experiments with two main systems: the CATT Encoder-Decoder model and the CATT-Whisper Early Fusion model. For both architectures, we first evaluated the official pre-trained checkpoints without any additional fine-tuning, followed by the application of our post-processing module. We then performed task-specific fine-tuning on the official KSAA-2026 training data.

5.1. Preprocessing

We followed the preprocessing pipeline used in the original CATT models (Alasmary et al., 2024), but set the Tashkeel-to-text ratio threshold to 0, retaining all samples due to the high quality and limited size of the KSAA-2026 training data.

5.2. Fine-tuning Setup

We fine-tuned both the CATT Encoder-Decoder and CATT-Whisper (Early Fusion) models on the official KSAA-2026 training split (2,327 utterances). Training was performed for a maximum of 20 epochs with early stopping based on validation DER. We used a batch size of 32 and a dropout rate of 0.01.

The AdamW optimizer was employed with layer-wise learning rates. We assigned a higher learning rate (1×10^{-3}) to the final classification head (and the speech projection head in CATT-Whisper) because these components are task-specific and benefit from faster adaptation. A much lower learning rate (4×10^{-5}) was used for the last layer of the encoder/decoder to gently update the pretrained knowledge without catastrophic forgetting. All remaining layers were frozen to preserve the rich representations learned during pretraining and to reduce computational cost and overfitting risk on the relatively small KSAA training set.

5.3. Post-processing

To mitigate the tendency of both CATT and CATT-Whisper models to drop or distort non-Arabic tokens (punctuation, numbers, symbols, and whitespace), we apply a deterministic post-processing step. The module preserves all formatting and non-Arabic characters from the original transcript while retaining the model's predicted diacritics on Arabic letters. Specifically, newlines and whitespace are taken exclusively from the original input, Arabic base letters and their diacritics are taken from the model output, and all other characters are copied directly from the original. An illustrative example is shown in Table 2. This alignment-based restoration significantly improves consistency and final evaluation scores.

Example Input	وهي قد تخرج خطاطين (متقنين)..
CATT-Whisper Output	وَهِيَ قَدْ تَخْرُجُ خَطَاطِينَ مُتْقِنِينَ
After Post-Processing	وَهِيَ قَدْ تَخْرُجُ خُطَّاطِينَ (مُتْقِنِينَ)..
Reference	وَهِيَ قَدْ تُخْرِجُ خَطَّاطِينَ (مُتْقِنِينَ)..

Table 2: Example from the KSAA-2026 development/test set showing CATT-Whisper output before and after applying the post-processing step, aligned with the reference.

6. Results

We evaluated our systems on both the development and hidden test sets of the KSAA-2026 Shared Task (Subtask 2). Tables 3, 4, and 5 present the Diacritic Error Rate (DER), Word Error Rate (WER), and Sentence Error Rate (SER) on the development set, respectively. Tables 6, 7, and 8 show the corresponding results on the hidden test set. All results are reported under two evaluation settings: with and without case-ending diacritics, and with and without including characters labeled as "no diacritic" in the evaluation.

In the zero-shot setting, CATT and CATT-Whisper achieve relatively similar performance. However, after fine-tuning on the KSAA-2026 training data, CATT-Whisper consistently outperforms CATT, particularly in WER (28.56 vs 33.50 on the development set). This demonstrates the benefit of combining acoustic features with task-specific adaptation.

On the hidden test set, our fine-tuned CATT-Whisper achieves the best performance with a DER of 7.04, WER of 24.39, and SER of 71.65. It outperforms all baselines, including the Fine-tuned Text+ASR baseline, confirming the advantage of multimodal early fusion. The post-processing step further improved the results by recovering punctuation and non-Arabic tokens frequently omitted by the base models.

7. Conclusion

In this paper, we presented our submission to the KSAA-2026 Shared Task (Subtask 2). We adapted the CATT-Whisper Early Fusion architecture and applied selective fine-tuning with differential learning rates on the official training data, followed by a post-processing module to restore non-Arabic tokens and punctuation. On the hidden test set, our best model achieved a DER of 7.04, WER of 24.39, and SER of 71.65, securing second place in the competition. These results demonstrate the effectiveness of multimodal approaches for speech-aware Arabic diacritization and highlight the importance of task-specific fine-tuning and output refinement.

Model	Including 'no diacritic'		Excluding 'no diacritic'	
	w. CE	w.o CE	w. CE	w.o CE
CATT (zero-shot)	13.08	7.02	15.62	8.07
Fine-tuned CATT	9.63	5.78	11.38	6.51
CATT-Whisper (zero-shot)	13.15	7.51	15.91	8.86
Fine-tuned CATT-Whisper	**8.06**	**5.84**	**9.30**	**6.47**
Text-only Baseline	20.97	15.13	23.29	15.21

Table 3: Diacritic Error Rate (DER) on the development set.

Model	Including 'no diacritic'		Excluding 'no diacritic'	
	w. CE	w.o CE	w. CE	w.o CE
CATT (zero-shot)	42.65	17.91	41.27	16.90
Fine-tuned CATT	33.50	**15.14**	32.04	**13.87**
CATT-Whisper (zero-shot)	42.95	19.55	41.79	18.58
Fine-tuned CATT-Whisper	**28.56**	16.07	**26.54**	14.21
Text-only Baseline	55.33	33.68	52.52	28.07

Table 4: Word Error Rate (WER) on the development set.

Model	Including 'no diacritic'		Excluding 'no diacritic'	
	w. CE	w.o CE	w. CE	w.o CE
CATT (zero-shot)	79.23	63.08	79.23	60.77
Fine-tuned CATT	**78.46**	**57.31**	**76.54**	**54.62**
CATT-Whisper (zero-shot)	82.31	68.46	82.31	66.92
Fine-tuned CATT-Whisper	81.15	63.46	79.23	58.46
Text-only Baseline	94.62	83.85	94.23	81.54

Table 5: Sentence Error Rate (SER) on the development set.

Model	Including 'no diacritic'		Excluding 'no diacritic'	
	w. CE	w.o CE	w. CE	w.o CE
Text-only Baseline	17.66	13.23	20.08	13.93
Text+ASR Baseline	13.50	10.58	14.26	9.96
Fine-tuned Text+ASR Baseline	9.91	7.89	8.52	**4.82**
Our Fine-tuned CATT-Whisper	**7.04**	**6.17**	**6.92**	5.15

Table 6: Diacritic Error Rate (DER) on the hidden test set.

Model	Including 'no diacritic'		Excluding 'no diacritic'	
	w. CE	w.o CE	w. CE	w.o CE
Text-only Baseline	49.85	32.24	46.20	27.07
Text+ASR Baseline	40.24	27.95	33.03	19.71
Fine-tuned Text+ASR Baseline	31.84	20.99	24.73	**10.89**
Our Fine-tuned CATT-Whisper	**24.39**	**17.96**	**18.97**	11.06

Table 7: Word Error Rate (WER) on the hidden test set.

Model	Including 'no diacritic'		Excluding 'no diacritic'	
	w. CE	w.o CE	w. CE	w.o CE
Text-only Baseline	91.77	82.62	91.77	81.71
Text+ASR Baseline	82.32	71.95	75.61	60.37
Fine-tuned Text+ASR Baseline	82.93	67.07	78.66	**50.61**
Our Fine-tuned CATT-Whisper	**71.65**	**55.18**	**68.90**	**50.61**

Table 8: Sentence Error Rate (SER) on the hidden test set.

8. Bibliographical References

Abdulmohsen Al-thubaity, Atheer AlKhalifa, Abdulrahman Almuhareb, and Waleed Alsanie. 2020. Arabic diacritization using bidirectional long short-term memory neural networks with conditional random fields. *IEEE Access*, PP:1–1.

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. pages 361–365.

Ahmad Ghannam, Naif Alharthi, Faris Alasmary, Kholood Al Tabash, Shouq Sadah, and Lahouari Ghouti. 2025. Abjad AI at NADI 2025: CATT-whisper: Multimodal diacritic restoration using text and speech representations. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 757–761, Suzhou, China. Association for Computational Linguistics.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–326, Abu Dhabi, UAE. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sara Shatnawi, Sawsan Alqahtani, and Hanan Aldarmaki. 2024. 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.

Dimitra Vergyri and Katrin Kirchhoff. 2004. Automatic diacritization of Arabic for acoustic modeling in speech recognition. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 66–73, Geneva, Switzerland. COLING.