Thaka at KSAA-2026 Task 2: Regularized Fine-Tuning for Arabic Speech Diacritization

Meshal Alamr, Hassan Alqaeri, Abdullah Aldahlawi
Thaka, Advanced AI and Information Technology
Riyadh, Saudi Arabia
{m.alamr, h.alqaeril, aldahlawi}@thakaait.net

Abstract

We describe the winning system for Task 2 of the KSAA-2026 Shared Task on Arabic Speech Dictation with Automatic Diacritization (Al Wazrah et al., 2026). The task requires producing fully diacritized Arabic text from speech audio and undiacritized transcripts, with only 2,327 training samples available and no external data permitted. Our system fine-tunes CATT-Whisper (Ghannam et al., 2025), a character-level multimodal model combining a pretrained CATT text encoder with a frozen Whisper speech encoder. The key to our approach is training regularization: R-Drop consistency regularization, Optuna-optimized hyperparameters with high weight decay, and Focal Loss. At inference, we average 200 stochastic forward passes across four model checkpoints using Monte Carlo Dropout at the softmax probability level. The system achieves **23.26% WER** on the primary leaderboard metric (with case endings, including no-diacritic positions), placing 1st among all participants.

Keywords: Arabic diacritization, multimodal, speech processing, regularization, shared task, KSAA-2026

1. Introduction

Arabic text is typically written without diacritics, the short vowels and phonological markers that determine pronunciation and meaning. Restoring these marks automatically is important for text-to-speech, machine translation, and language learning (Habash and Rambow, 2007), but remains challenging: text-only models (Alasmary et al., 2024) struggle with inherent ambiguity, particularly in dialectal Arabic where phonetic variation is high. Speech signals offer a complementary source of disambiguation, as prosodic and phonetic cues can resolve cases that are irrecoverable from orthography alone.

The KSAA-2026 Shared Task (Al Wazrah et al., 2026) presents a low-resource instance of this problem: only 2,327 training samples across multiple Arabic dialects, with no external data permitted. CATT-Whisper (Ghannam et al., 2025) demonstrated that a multimodal architecture combining CATT and Whisper encoders is effective for this task, achieving strong results at NADI 2025. We adopt this architecture and focus on making the most of the limited training data through regularization and inference-time ensembling.

2. Background

2.1. Task and Data

Task 2 of the KSAA-2026 Shared Task requires participants to produce fully diacritized Arabic text given speech audio and undiacritized transcripts. Systems are evaluated using Diacritic Error Rate (DER), Word Error Rate (WER), and Sentence Error Rate (SER) (Fadel et al., 2019), with **WER** (with case endings, including no-diacritic positions) as the primary ranking metric. All results in this paper are reported under this setting.

The dataset (Al Wazrah et al., 2026) contains 2,327 training, 260 development, and 328 test samples spanning multiple Arabic dialects. Each sample is a short utterance (average $\sim$7 seconds) paired with a diacritized transcript. We filter samples where the ratio of diacritized to total characters is below 0.6, yielding 2,187 effective training samples.

2.2. Related Work

Arabic diacritization has progressed from morphological taggers (Habash and Rambow, 2007) through transformer-based systems (Nazih and Hifny, 2022) to pretrained encoders such as AraBERT (Antoun et al., 2020) and the character-level CATT model (Alasmary et al., 2024). Ghannam et al. (2025) introduced CATT-Whisper, which fuses a CATT text encoder with Whisper (Radford et al., 2023) speech features via prefix addition, winning the NADI 2025 dialectal diacritization track. R-Drop (Liang et al., 2021) penalizes divergent predictions between two dropout-masked forward passes and has been effective in low-resource settings. MC Dropout (Gal and Ghahramani, 2016) enables ensemble-like inference from a single model by keeping dropout active at test time. Our work applies these regularization and inference techniques to multimodal Arabic diacritization.

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 225–228
May 11, 2026.

Hyperparameter	Value
Learning rate	4.1×10^{-6}
R-Drop α	2.08
Focal γ / label smoothing	0.34 / 0.018
Weight decay	0.098
Speech emb. dropout	0.09
Batch size / epochs	16 / 40
Warmup / min LR factor	3 ep / 0.002
SpecAugment (freq / time)	10 / 63
Noise injection	Gauss., SNR 10–30 dB
Whisper unfrozen	0 (fully frozen)

Table 1: Hyperparameters selected via Optuna (30 trials).

3. System Overview

3.1. Architecture

Our system builds on CATT-Whisper. The text encoder is a 6-layer CATT Transformer ($d = 512$, 16 heads), pretrained on Arabic diacritization, predicting one of 15 diacritic classes per Arabic letter. The speech encoder is Whisper-base (Radford et al., 2023) (6 encoder blocks, $d = 512$), kept fully frozen in the primary configuration. Fusion uses prefix addition: 1,500 Whisper frames are mean-pooled into 150 tokens, projected, and added to 150 dedicated prefix positions that precede the text input. The model has $\sim$39M parameters ($\sim$19M trainable).

3.2. Training

Figure 1(a) illustrates the training procedure. We fine-tune with R-Drop: each input passes through the model twice with different dropout masks, and a symmetric KL divergence penalty ($\alpha = 2.08$) encourages consistent predictions. The supervised loss is Focal Loss (Lin et al., 2017) ($\gamma = 0.34$) with label smoothing ($\epsilon = 0.018$). Speech embedding dropout ($p = 0.09$) randomly zeros the speech representation during training, following the modality-robust training scheme of CATT-Whisper. Audio augmentation uses SpecAugment (Park et al., 2019) and Gaussian noise injection. Hyperparameters were selected using Optuna (Akiba et al., 2019) (30 trials, 12 epochs each); Table 1 shows the final configuration. We use AdamW (Loshchilov and Hutter, 2019) with cosine learning rate decay. We train four checkpoints: three with this configuration using seeds 42, 7, and 123, and a fourth from a separate Optuna trial with an alternative configuration (learning rate 4.7×10^{-5}, batch size 32, $\gamma = 1.0$, label smoothing 0.108, and 4 unfrozen Whisper blocks after epoch 15) to increase ensemble diversity.

System	DER↓	WER↓	SER↓
meshal (Ours)	**6.87**	**23.26**	**66.16**
nadaadelmousa	7.04	24.39	71.65
naif_alharthi	7.51	25.34	73.48
nahian_abu	8.23	30.37	80.79
Hassan	10.56	34.47	79.88
omarnj10	27.94	44.05	98.78
astral_fate	31.67	84.50	99.70
Baseline (FT text+ASR)	9.91	31.84	82.93
Baseline (text+ASR)	13.50	40.24	82.32
Baseline (text-only)	17.66	49.85	91.77

Table 2: Test set results. All metrics: with case endings, incl. no-diacritic positions. Ranked by WER (primary metric).

3.3. Inference

Figure 1(b) illustrates the inference procedure. At inference, we keep dropout active in the CATT encoder ($p = 0.1$) while LayerNorm stays in eval mode. Each of the four models runs 50 stochastic forward passes ($4 \times 50 = 200$ total), and we average softmax probabilities across all passes before taking the argmax.

3.4. Post-Processing

We use direct positional diacritic insertion: CATT models maintain 1:1 correspondence between Arabic letter positions and predicted diacritics. We enforce three invariants: (1) stripping diacritics from the output recovers the original input; (2) diacritic count matches predictions; (3) all letter positions are consumed.

4. Results

Table 2 shows the test set results compared to other participants and the shared task baselines. Our system achieves **23.26% WER** on the primary metric.

Table 3 presents a cumulative ablation on the development set, starting from the pretrained CATT-Whisper model and progressively adding our modifications. Fine-tuning CATT-Whisper with a standard recipe (cross-entropy loss, learning rate 10^{-5}) gives 30.43% WER; adding our regularized recipe (R-Drop, Focal Loss, high weight decay) reduces this to 27.18%, and MC Dropout ensembling brings it to 26.02%. The majority of the gain (3.25 pp) comes from the regularized training recipe; MC Dropout adds a further 1.16 pp, confirming that the training recipe is the primary driver of improvement.

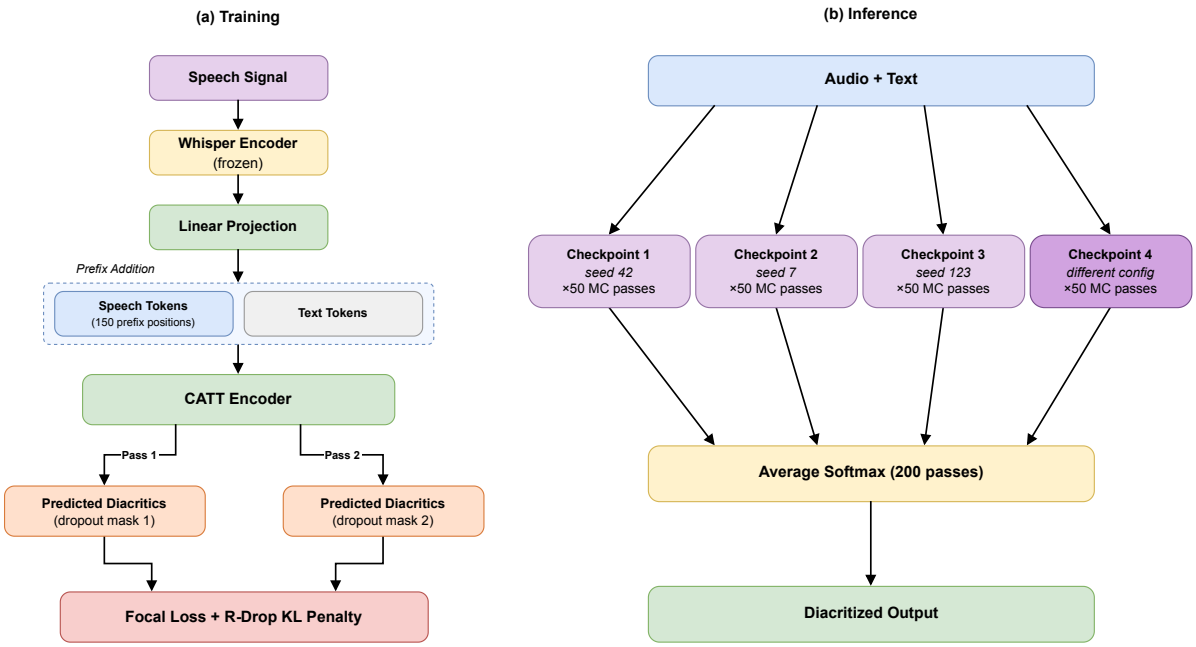

Figure 1: (a) Training: speech features from the frozen Whisper encoder are fused with text tokens via prefix addition and processed by the CATT encoder. R-Drop runs two forward passes with different dropout masks, optimizing with Focal Loss and a KL consistency penalty. (b) Inference: four checkpoints each run 50 MC Dropout passes; the 200 softmax distributions are averaged to produce the diacritized output.

Configuration	DER	WER
CATT-Whisper (pretrained)	17.76	54.06
CATT-Whisper (fine-tuned)†	8.59	30.43
+ Regularized recipe‡	7.57	27.18
+ 4-model MC Dropout ensemble	**7.17**	**26.02**

†Baseline: lr=10^{-5}, cross-entropy loss, batch size 16, 30 epochs.

‡R-Drop + Focal Loss + high weight decay (Optuna, 30 trials).

Table 3: Cumulative ablation on dev set (%, with case endings, incl. no-diacritic positions).

Input	الظاهر أنه لا خلاف في الحقيقة للاتفاق على امتناع إدراك حقيقة الذات
Ours	الظَّاهِرُ أَنَّهُ لَا خِلَافَ فِي الْحَقِيقَةِ لِلِاتِّفَاقِ عَلَى امْتِنَاعِ إِدْرَاكِ حَقِيقَةِ الذَّاتِ
Gold	الظَّاهِرُ أَنَّهُ لَا خِلَافَ فِي الْحَقِيقَةِ لِلِاتِّفَاقِ عَلَى امْتِنَاعِ إِدْرَاكِ حَقِيقَةِ الذَّاتِ

Table 4: Dev set example.

4.1. Discussion

Training recipe. We also explored architectural modifications including cross-attention fusion, CRF decoding, attention pooling, auxiliary heads, and RL fine-tuning, none of which improved over the fine-tuned CATT-Whisper baseline. The regularized recipe yielded a 3.25 pp WER gain (Table 3), suggesting that, in our experiments with 2,187 training samples, the optimization strategy matters more than the model architecture. MC Dropout ensembling adds 1.16 pp at a cost of 200 forward passes ($\sim$50$\times$ slower than a single pass); in practice, reducing the number of passes and studying the effect on accuracy could yield a more efficient trade-off.

Audio contribution. Ghannam et al. (2025) showed that incorporating speech features significantly improves diacritization accuracy. Consistent with their findings and the gap between the text-only and audio-equipped baselines in Table 2, we observed a similar pattern when fine-tuning CATT-Whisper without speech features, further confirming the importance of audio input for this task.

Qualitative example. Table 4 shows a dev set example where our system correctly diacritizes a complex sentence with case endings and shaddas.

5. Conclusion

We presented our system for KSAA-2026 Task 2, which fine-tunes CATT-Whisper with R-Drop, Focal Loss, and Optuna-optimized hyperparameters,

and uses MC Dropout ensembling at inference. At this data scale, training regularization yielded larger gains than any of the architectural modifications we explored. Future work could disentangle the individual contributions of each regularization component and examine per-dialect performance to identify remaining challenges.

Acknowledgements

We thank Thaka for supporting this work, KSAA for organizing the shared task, and the Abjad AI team for open-sourcing the CATT-Whisper model.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

Asma Al Wazrah, Waad Alshammari, Rawan Almatham, Raghad Al-Rasheed, Afrah Altamimi, Rufael Marew, Sawsan Alqahtani, Hanan Aldarmaki, Abdullah Alharbi, Abdulrahman Alshehri, Mohamed Assar, Amal Almazrua, and Abdulrahman AlOsaimy. 2026. KSAA-2026 shared task on Arabic speech dictation with automatic diacritization. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7)*.

Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. 2024. CATT: Character-based Arabic tashkeel transformer. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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 (OSACT)*, pages 9–15.

Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. 2019. Arabic text diacritization using deep neural networks. *arXiv preprint arXiv:1905.01965*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059.

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 Second Arabic Natural Language Processing Conference: Shared Tasks (ArabicNLP 2025)*.

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.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Waleed Nazih and Yasser Hifny. 2022. Arabic syntactic diacritics restoration using BERT models. *Computational Intelligence and Neuroscience*, 2022:3214255.

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 *Interspeech*, 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 (ICML)*, pages 28492–28518.