Riwaya-ID: Towards ML-powered Identification of Qur'anic Recitation Style from Audio

A. Anas Chentouf
Massachusetts Institute of Technology
chentouf@mit.edu

Abstract

The Holy Qur'an, the scripture of Muslims, is recited through slightly different transmission traditions (riwayat) encoding different recitation rules. In this paper, we study riwaya identification: determining the Qur'anic transmission style directly from audio. In order to do so, we curate over 700 hours of recitations and segment recordings into 12 s windows to build a dataset. Building on pretrained speech encoders (e.g., `wav2vec2.0`, `Whisper`), we extract frame-level embeddings and train a classifier to predict the riwaya. Our embedding-based models achieve an 82% prediction accuracy in distinguishing Warsh from Hafs. We hope that this work leads to scalable, audio-native tools supporting different recitation styles using modern pretrained encoders.

1 Introduction

The task of Spoken Language Identification (SLI) is often studied as a building block that enables large-scale speech-capable intelligent systems to serve a diverse set of users. A more delicate task is Spoken Dialect Identification (SDI), where the goal is to identify which dialect of the language is being spoken given an audio clip. This slight change introduces an additional layer of complexity, particularly in low-resource languages or dialects Lonergan et al. [2023].

Beyond general speech applications, the tools of SLI and SDI can be applied to specialized domains where subtle differences in vocabulary or pronunciation occur. One such domain is Qur'anic recitation, where different riwayat[1] represent authentic transmission chains, and where great emphasis is placed on correct pronunciation. Identifying these automatically from audio would open the door to richer digital archives and more inclusive recitation tools.

Table 1: Examples of differences between Hafs and Warsh, the two most prominent riwayat.

Verse	**Hafs Text**	**Warsh Text**	**Note**
Al-Fātiḥa 1:4	*Mālik*	*Malik*	Vowel length; Owner vs King.
Al-Baqarah 2:125[2]	*wa-ttakhidhu*	*wa-ttakhadhu*	Morphology: imperative vs past.
Al-Mu'minūn 23:1	*al-mu'minūn*	*al-muminūn*	Pronunciation (hamza).

While recent years have witnessed various efforts in digitizing Qur'anic resources, ranging from font libraries to verse-by-verse segmented audio, the vast majority of such resources are devoted exclusively to the *riwaya* of Hafs 'an 'Asim. A *riwaya*, or transmission, is an authentic method of reciting the Qur'an that is transmitted from the Prophet PBUH. Different riwayat have slightly different recitation rules (*aḥkām*). The task of detecting a riwaya from audio would be beneficial in labeling audio clips to expand Qur'anic digital libraries, serving Muslims who recite in other riwayat.

[1] plural of riwaya

Similarly, riwaya detection would enable expanding AI-powered Qur'anic tools to other riwayat. Recently, models have been trained to transcribe, segment [Abdelfattah et al., 2025], correct [Alagrami and Eljazzar, 2020], and even match Qur'anic recitation to text [Tall et al., 2023]. A prominent example is `tarteel.ai`, an AI-powered memorization app, which uses Automatic Speech Recognition (ASR) models to transcribe and correct live Qur'anic recitation. As it currently stands, both Qur'anic audio models and datasets are usually in Hafs, a limitation which we attempt to tackle in this paper.

Contributions The contributions of this paper can be summarized as follows: we **(1)** introduce *Riwaya-ID*, formally defining the task and framework for identifying the *riwaya*; **(2)** curate a weakly labeled corpus of Qur'anic recitations; and **(3)** establish baselines for this task. We expect to release the data in the upcoming preprint version.

1.1 Related Works

The approach of learning semantically meaningful embeddings underlies many recent artificial intelligence (AI) systems, across various modalities. Pretrained models capture rich representations that can be finetuned for downstream tasks. For example, `word2vec` [Mikolov et al., 2013] encodes word meaning from context, while `wav2vec` [Schneider et al., 2019] learns audio representations by relating short units ($\approx$ 25 ms) to surrounding context. Its successor, `wav2vec 2.0` [Baevski et al., 2020], replaces convolutional context networks with Transformers, combining local patterns with global dependencies. OpenAI's `Whisper` [Radford et al., 2022], by contrast, converts raw audio into log-Mel spectrograms, which are closely related to Mel-Frequency Cepstral Coefficients (MFCCs), and processes them with a Transformer encoder–decoder.

Earlier SLI methods relied on MFCCs, pitch, and similar cues [O'Shaughnessy, 2025]. Neural approaches extended this: convolutional networks learned features from Mel-spectrograms [Singh et al., 2021], while time-delayed networks modeled temporal dynamics in the audio [Kepecs and Beigi, 2022]. Similar techniques have been applied to dialect classification, especially in low-resource settings [Lonergan et al., 2023, Das et al., 2023, Fischbach et al., 2025].

For Qur'anic recitation, generic Arabic ASR is insufficient due to the unique *aḥkām* (rules) governing length, stress, and intonation, which differ from both Modern Standard Arabic and colloquial dialects. Research has addressed these challenges through pronunciation benchmarking [Kheir et al., 2025], phoneme-level transcription difficulties [Zaatiti et al., 2025], and rule-specific ASR systems [Alagrami and Eljazzar, 2020]. Previous works have also tackled riwaya detection using MFCC and Hidden Markov Models Yousfi and Zeki [2016], Das et al. [2023], often focusing on the *madd* (elongation). We highlight a subset of Qur'an-specific ASR models below.

Table 2: Recent Qur'anic speech models and their base architectures, from Hugging Face.

Model	**Base Model**	**Year**
`IbrahimSalah/Wav2vecLarge-quran-syllables`	`facebook/wav2vec2-large`	2024
`tarteel-ai/whisper-base-ar-quran`	`openai/whisper-base`	2022
`tarteel-ai/whisper-tiny-ar-quran`	`openai/whisper-tiny`	2022

2 Methods

Let $\mathcal{X}$ denote the input feature space, representing the space of recitation (audio clips). Let $\mathcal{Y} = \{1, \cdots, R\}$ denote the output label space of riwayat. Our architecture uses pretrained speech models to obtain meaningful audio embeddings. Formally, an embedding model $f_\theta : \mathcal{X} \to \mathbb{R}^d$ maps each input audio segment $x \in \mathcal{X}$ into a latent embedding space. In our case, f_θ is instantiated using pretrained models such as `wav2vec2.0` or `Whisper` and their variants which were pretrained on Qur'anic audio, as described in Table 2. The obtained embedding is contextualized at the frame-level.

To adapt the pretrained backbone to riwaya identification, we append a lightweight, feed-forward classification head $g_\phi : \mathbb{R}^d \to \mathbb{R}^R$ with one hidden layers.

Table 3: The embedding-based classifier results are reported as mean ± std, over 3 random seeds. Base, generic models are noted with †, while the other models are finetuned on Arabic or Qur'anic audio. The base models are trained on the same hyperparameters as the corresponding Qur'an models.

Author	Base Model	Test Acc.	Logit Mean Agg.	Max Agg.	Prob. Mean Agg.	Maj. Agg.
tarteel-ai	whisper-base	80.79 ± 0.26	**83.69 ± 0.63**	**85.37 ± 0.69**	**83.21 ± 0.69**	**82.53 ± 0.83**
IbrahimSalah	wav2vec2-large	79.87 ± 0.16	78.67± 0.27	79.98± 0.76	79.50± 0.54	79.42± 0.39
jonatasgrosman	wav2vec2-large-xlsr-53	78.14 ± 0.14	79.79± 0.65	80.04± 0.38	79.57± 0.63	79.14± 0.77
openai†	whisper-base	**81.45 ± 0.23**	82.60 ± 0.59	82.74 ± 0.92	82.00 ± 0.41	81.89 ± 0.71
facebook†	wav2vec2-large	54.74 ± 3.79	56.48 ± 5.63	57.31 ± 4.60	56.52 ± 5.65	56.45 ± 5.43

2.1 Training

We train our model by minimizing the cross-entropy loss between predicted distributions and assigned labels. During training, the backbone is initially frozen to allow the classification head to learn a meaningful representation, before unfreezing the layers progressively. Using the `AdamW` optimizer [Loshchilov and Hutter, 2017], we specify different learning rates for the backbone and the classification head. Moreover, we employ Layer-wide Learning Rate Decay (LLRD) Bao et al. [2021] to decrease the learning rate for earlier transformer layers. For more details on our data collection and experiment details, see Appendix A and B, resp.

For the purposes of model training, we mostly restrict the task to binary classification between *Hafs 'an 'Asim* and *Warsh 'an Nafi'*. This choice is motivated by several factors: **(1)** computational constraints make it impractical to train on all available riwayat, **(2)** these two classes are the most reliable in terms of label quality, and are also by far the most common riwayat, and **(3)** the performance of our methodology on multiclass classification (6 riwayat) was subpar.

We draw data from our corpus so that classes are represented equally in training and test. We also ensure segments derived from one audio recording don't leak across training and non-training splits.

3 Results

We evaluate riwaya identification on the test split, reporting the classification accuracy. For embedding models, we perform an extensive hyperparameter sweep (optimizer, learning rate, weight decay, dropout, classifier depth, and pooling) and report the test accuracy corresponding to the **best** validation accuracy. We also report the accuracy obtained by aggregating across the entire audio clip, experimenting with three aggregation strategies: averaging the logits (Logit Mean), consider the most confident logit across the window (Max), averaging the softmax probabilities (Prob. Mean), and majority voting (Maj. Ag). See Appendix D for an explanation of the aggregation strategies.

Baselines. As baselines, we explore the use of embedding models which have **not** been specifically fine-tuned for Qur'anic audio. Frame-level representations are aggregated via mean pooling. We also evaluate against `jonatasgrosman`'s `wav2vec2-large-xlsr-53` model which has been fine-tuned on *general* Arabic (and not Qur'anic) audio.

The given set of `Whisper` models outperforms the `wav2vec-large`. Interestingly, the `Whisper` base model *outperforms* the `tarteel-ai` model in classifying the riwaya from 12 seconds of audio. However, the `wav2vec2` models which are further trained on Arabic or Qur'anic data perform significantly better than the base `wav2vec2` model. This changes when we allow our model to aggregate over the entire recording: `tarteel-ai/whisper-base-quran` outperforms the base model by a significant margin.

In general, aggregating improves model performance as it allows our model to depend less on a particular window and more on the training data (which includes the Surah being recited). One can probabilistically justify this, e.g. via Chernoff or Hoeffding bounds. Among the various aggregation strategies, we observe that maximum logit aggregation strategy works best: using the model's most confident guess at any point during the entire recording, usually a subset of the *surah*, or chapter. Across all aggregation strategies, `tarteel-ai` models achieve the highest accuracy.

Another observation we make is that our model is more likely to predict Hafs for Warsh clips than predict Warsh for Hafs. A possible cause for this is that the models we use were actually finetuned on Hafs. See Figure 1 for confusion matrices from our experiments.

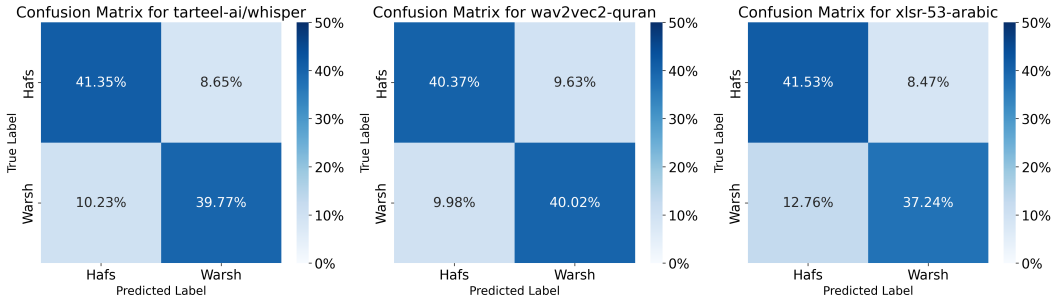

Figure 1: Confusion matrix for `tarteel-ai/whisper-base-quran-ar` (left), `IbrahimSalah/Wav2vecLarge_quran_syllables_recognition` (middle), and `jonatasgrosman/wav2vec2-large-xlsr-53-arabic` (right) on binary classification.

We also experiment with the `tarteel-ai/whisper-base` model trained on an *incomplete* list of six riwayat: Hafs 'an 'Asim, Warsh 'an Nafi', Qalun 'an Nafi, Al-Bazzi 'an Ibn Kathir, Qunbul 'an Ibn Kathir, Ibn Jummaz 'an Abu Ja'far. This model obtains a lower test accuracy of 62.50%: demonstrating that our approach needs refining. The confusion matrix can be found in Appendix C.

4 Conclusion

Qur'anic recitation is central to Islamic practice, and its diverse riwayat reflect a rich tradition of transmission. Our work provides a proof-of-concept approach identifying these recitations from audio, with the aim of enabling more inclusive digital resources and tools. By leveraging pretrained speech models such as `wav2vec2.0` and `Whisper`, we showed that even a simple classification head can distinguish between major riwayat with promising accuracy.

4.1 Limitations

Data. As noted earlier, our dataset is a major limitation. Labels are assigned via keyword searches, which introduces noise whose magnitude we cannot currently estimate. Additionally, some educational clips mix multiple riwayat, further complicating supervision. Our experiment focused primarily on binary classification, which is an extremely limited case.

Turuq We restricted our attention to identifying major riwayat (e.g., Hafs, Warsh, etc..), however, a riwaya may branch further into *ṭuruq* (transmission paths). Each of those *ṭuruq* encodes even finer rules of recitation, but the author is not aware of a digital annotated corpus of turuq audio.

4.2 Future Work

We outline a few directions which we hope to explore.

Predicting via aggregated logits. Logit aggregation across an entire recording prevents us from accurately applying riwaya detection at a streaming rate. One wonders how varying the window length affects the results; perhaps one can obtain similar performance gains by only aggregating logits across a smaller window. This would allow deploying the model on live recitations.

Repositories of riwayat. Human-annotated datasets of different riwayat serve as a two-way resource: they can be used both to train improved models and, conversely, to expand the datasets themselves through model-assisted annotation. As it currently stands, the author is not aware of large-scale digital databases of recitations involving different riwayat.

Label noise. Future work can incorporate techniques from the label-noise literature, such as generalized cross-entropy loss [Zhang and Sabuncu, 2018], early stopping [Yuan et al., 2023], or other robust training methods [Wei et al., 2021] that account for open noise.

Acknowledgements

The author would like to thank Shaden Alshammari, Habeeb Salau, Husam El-Nager, Mohamed Samb, and Ahmed Katary for their feedback and helpful discussions. The author would also like to thank the anonymous reviewers for their suggestions.

References

Abdullah Abdelfattah, Mahmoud I Khalil, and Hazem Abbas. Automatic pronunciation error detection and correction of the holy quran's learners using deep learning. *arXiv [eess.AS]*, August 2025.

Ali M Alagrami and Maged M Eljazzar. Smartajweed automatic recognition of arabic quranic recitation rules. *arXiv [eess.AS]*, December 2020.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv [cs.CL]*, June 2020.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv [cs.CV]*, June 2021.

Sourya Dipta Das, Yash Vadi, Abhishek Unnam, and Kuldeep Yadav. Unsupervised out-of-distribution dialect detection with mahalanobis distance. *arXiv [cs.CL]*, August 2023.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. Improving low-resource dialect classification using retrieval-based voice conversion. *arXiv [cs.CL]*, July 2025.

Benjamin Kepecs and Homayoon Beigi. Automatic spoken language identification using a time-delay neural network. *arXiv [cs.CL]*, May 2022.

Yassine El Kheir, Omnia Ibrahim, Amit Meghanani, Nada Almarwani, Hawau Olamide Toyin, Sadeen Alharbi, Modar Alfadly, Lamya Alkanhal, Ibrahim Selim, Shehab Elbatal, Salima Mdhaffar, Thomas Hain, Yasser Hifny, Mostafa Shahin, and Ahmed Ali. Towards a unified benchmark for arabic pronunciation assessment: Quranic recitation as case study. *arXiv [cs.SD]*, June 2025.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. Towards spoken dialect identification of irish. *arXiv [cs.CL]*, July 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv [cs.LG]*, November 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv [cs.CL]*, January 2013.

Douglas O'Shaughnessy. Spoken language identification: An overview of past and present research trends. *Speech Commun.*, 167(103167):103167, February 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and I Sutskever. Robust speech recognition via large-scale weak supervision. *ICML*, pages 28492–28518, December 2022.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv [cs.CL]*, April 2019.

Gundeep Singh, Sahil Sharma, Vijay Kumar, Manjit Kaur, Mohammed Baz, and Mehedi Masud. Spoken language identification using deep learning. *Comput. Intell. Neurosci.*, 2021(1):5123671, September 2021.

Mohamet Tall, T I Diop, Ndeye Fatou Ngom, and El Hadji Abdoulaye Thiam. Deep learning for quranic reciter recognition and audio content identification. *13th Conference on Research in Computer Science and its Applications (CNRIA), 25-27 May 2023*, May 2023.

Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. Open-set label noise can improve robustness against inherent label noise. *arXiv [cs.LG]*, June 2021.

Bilal Yousfi and Akram M Zeki. Holy qur'an speech recognition system distinguishing the type of recitation. In *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, pages 1–6. IEEE, July 2016.

Suqin Yuan, Lei Feng, and Tongliang Liu. Early stopping against label noise without validation data. In *The Twelfth International Conference on Learning Representations*, October 2023.

Hadi Zaatiti, Hatem Hajri, Osama Abdullah, and Nader Masmoudi. Towards stable AI systems for evaluating arabic pronunciations. *arXiv [cs.CL]*, August 2025.

Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv [cs.LG]*, May 2018.

A Data

We collect over 700 hours of Qur'anic recitation by scraping online audio sources for recitations of each *sūrah* (chapter) across different riwayat (transmission methods). Because large-scale human-annotated resources are limited, we rely on keyword matching and search criteria to label the data. This weak supervision introduces noise into our corpus, which we acknowledge as a key limitation to be addressed in future work.

Each recording is segmented into 12-second windows with a stride of 3 seconds. Additional loudness normalization allows for normalizing the volume. These overlapping sub-clips inherit the label of the parent recording. Shorter windows allow for streaming-like performance but decrease the amount of information available, creating a delicate tradeoff which we hope to explore. A further challenge arises from the fact that some recordings contain more than one riwaya, meaning that inherited labels are not always perfectly accurate.

In this work, we focus on six of the most prominent riwayat, shown in Table 4. These do not represent the full diversity of recitations that exist in the Islamic tradition, but they cover some of the most widely practiced transmissions for which we could find data for. Among them, *Hafs 'an 'Asim* is the overwhelmingly dominant recitation, reportedly accounting for around 95% of global practice.[3]

Table 4: Hours of audio collected per riwāya

Riwaya	**Hours of Audio**
Hafs 'an 'Asim	161.76
Warsh 'an Nafi'	147.37
Qalun 'an Nafi'	133.29
Al-Bazzi 'an Ibn Kathir	104.27
Qunbul 'an Ibn Kathir	106.74
Ibn Jummaz 'an Abu Ja'far	75.85

The train, validation, and test splits account for 80 %, 10%, and 10% of the data, respectively. As described in the paper, we ensure that no subsets of the same clip are present in both train and validation or train and test splits. This is done to prevent memorization and leakage.

[3]This figure is often cited informally but, to the best of our knowledge, has not been established through a rigorous scientific study. Estimating prevalence can be done in future work once an accurate classifier has been built.

B Experimental Details

All experiments were conducted on a single machine equipped with 8 NVIDIA A100 GPUs using PyTorch `DataParallel`. Models were initialized from Hugging Face checkpoints and trained with the `AdamW` optimizer. Throughout the experiment, we adopt a batch size of 32. To adapt the pretrained backbones, we first froze all layers and progressively unfroze the last N transformer layers according to the hyperparameter sweep. Layer-wise Learning Rate Decay (LLRD) was applied with decay factors ranging from 0.95 to 0.99.

We performed hyperparameter sweeps over classification head size, dropout rate, learning rates, weight decay, and warmup schedules. Three pooling strategies were considered: mean pooling, stats pooling, and attention pooling. For each pooling method, we varied hidden dimension sizes, dropout, backbone vs. head learning rates, the number of unfrozen layers, and LLRD values. Additional sweeps included weight decay (0.005–0.02), freeze warmup steps (1–3), and warmup epochs (1–3). The best configuration for each model was selected on the validation set, with results reported in Table 3.

C Multiclass Experiment

The following is a confusion matrix obtained while training `tarteel-ai/whisper-base` to distinguish all available riwayat.

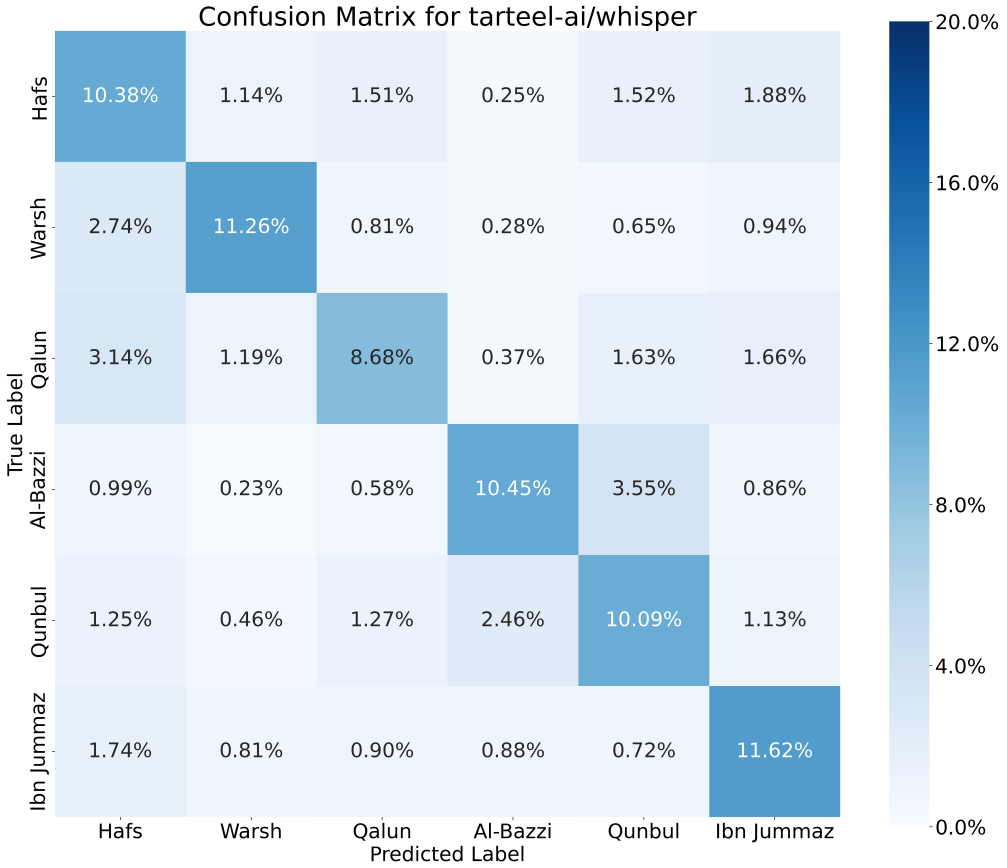

Figure 2: Confusion matrix for a `tarteel-ai/whisper-base` model which was trained on all 6 riwayat.

We particularly note some interesting features from this analysis. The confusion matrix demonstrates a relatively higher error between more similar riwayat that are narrated from the same scholar. In our dataset's case, Warsh and Qalun are both narrated from Imam Nafi', while Al-Bazzi and Qunbul are narated from Ibn Kathir. These pairs of riwayat share many recitation rules and hence, this phenomenon is not surprising.

D Logit Aggregation Strategies

General setup and notation. Let a sequence (e.g., all clips from one section/video) be indexed by time $t = 1, 2, \ldots, T$. For each clip t, the model outputs logits $z_t \in \mathbb{R}^R$ over R classes (riwayat, and probabilities

$$p_t \;=\; \mathrm{softmax}(z_t), \qquad (p_t)_c \;=\; \frac{e^{z_{t,c}}}{\sum_{j=1}^{R} e^{z_{t,j}}}.$$

Given a window length $n \in \mathbb{N}$ and a window $W_t = \{t, t+1, \ldots, t+n-1\}$ restricted to the same section/video, we aggregate $\{z_k, p_k\}_{k \in W_t}$ into a single distribution $\tilde{p}_t^{(n)} \in \Delta^{R-1}$ (the probability simplex). Final prediction is $\hat{y}_t^{(n)} = \arg\max_{c \in \{1,\ldots,R\}} \tilde{p}_{t,c}^{(n)}$. When $n = 1$ we recover per-clip predictions.

- **Logit Mean Aggregation.**

$$\tilde{p}_t^{(n)} \;=\; \mathrm{softmax}\Big(\frac{1}{n}\sum_{k \in W_t} z_k\Big), \qquad \tilde{p}_{t,c}^{(n)} \;\propto\; \exp\Big(\frac{1}{n}\sum_{k \in W_t} z_{k,c}\Big).$$

  Equivalently, this is the normalized geometric mean of the unnormalized scores; it tends to sharpen agreement across clips.

- **Max of probabilities.**

$$m_c \;=\; \max_{k \in W_t} (p_k)_c, \qquad \tilde{p}_{t,c}^{(n)} \;=\; \frac{m_c}{\sum_{j=1}^{R} m_j}.$$

  Element-wise maximum followed by renormalization (for a valid distribution); emphasizes any high-confidence evidence.

- **Probability Mean.**

$$\tilde{p}_t^{(n)} \;=\; \frac{1}{n}\sum_{k \in W_t} p_k, \qquad \tilde{p}_{t,c}^{(n)} \;=\; \frac{1}{n}\sum_{k \in W_t} (p_k)_c.$$

  This averages calibrated posteriors; it is smoothing and robust to scale mismatch across clips.

- **Majority vote.**

$$v_k \;=\; \arg\max_c (p_{k,c}), \qquad \tilde{p}_{t,c}^{(n)} \;=\; \frac{1}{n}\sum_{k \in W_t} \mathbf{1}\{v_k = c\}, \qquad \hat{y}_t^{(n)} \;=\; \arg\max_c \tilde{p}_{t,c}^{(n)}.$$

Remarks. All windows W_t are confined within a single section/video. One can easily verify that for binary classification with $R = 2$ clips and a window length of $n = 2$, the first and third strategies coincide. With binary classification, it suffices to consider a logit which is converted to a softmax via the sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$.

This amounts to showing that

$$\sigma\left(\frac{z_1 + z_2}{2}\right) > \frac{1}{2} \Leftrightarrow \frac{\sigma(z_1) + \sigma(z_2)}{2} > \frac{1}{2}.$$

Equivalently, one must show that

$$\frac{1}{1 + e^{-(z_1+z_2)/2}} > \frac{1}{2} \Leftrightarrow \frac{1}{1+e^{-z_1}} + \frac{1}{1+e^{-z_2}} > 1.$$

Expanding and rearranging, we get that the problem is equivalent to showing that

$$1 > e^{-(z_1+z_2)/2} \Leftrightarrow 1 > e^{-z_1} e^{-z_2},$$

which clearly holds.

However, for $n \geq 3$ the strategies need not agree. For example,

$$z_1 = [4.447, 1.723], \quad z_2 = [4.799, 1.254], \quad z_3 = [0.086, 6.752],$$

yield margins $m_1 = 2.724$, $m_2 = 3.545$, $m_3 = -6.666$. Then

$$\frac{\sigma(m_1) + \sigma(m_2) + \sigma(m_3)}{3} \approx 0.637 > 0.5 \quad \text{while} \quad \sigma\Big(\frac{m_1 + m_2 + m_3}{3}\Big) \approx 0.467 < 0.5,$$

so the two strategies predict different classes.

E Manually Curated Data

We evaluate our finetuned `tarteel-ai/whisper-base` model on a manually curated dataset of differences between Hafs and Warsh, both recited by Sh. Abdul Basit Abdul Samad. This data was obtained from `everyayah.com`. We use this to examine the model's predictions and identify particular failure/success modes.

Verse [4]	**Hafs**		**Warsh**	
	Mean $\hat{p}$(Hafs)	**Maj. Pred.**	**Mean $\hat{p}$(Warsh)**	**Maj. Pred.**
1:4	0.330	Warsh	0.624	Warsh
2:125	0.278	Warsh	0.734	Warsh
12:35	0.587	Hafs	0.680	Warsh
12:101	0.503	Hafs	0.672	Warsh
18:2	0.483	Hafs	0.964	Warsh
20:1	0.403	Warsh	0.598	Warsh
23:1	0.312	Warsh	0.628	Warsh
27:31	0.591	Hafs	0.501	Warsh
36:10	0.705	Hafs	0.444	Hafs
48:17	0.408	Warsh	0.795	Warsh
72:7	0.562	Hafs	0.764	Warsh
75:37	0.884	Hafs	0.570	Warsh
91:15	0.347	Warsh	0.597	Warsh
98:6	0.047	Warsh	0.913	Warsh

Table 5: Comparison of Abdul Basit Abdul Samad's recitations of the same verse in Hafs vs Warsh. We report the predicted confidence (as the probability of the relevant class) as well as the majority prediction taken over windows from the clip. We note that the v

One observes that Warsh is much accurately predicted on its class; whereas Hafs features a more confused model. A few of these feature instances where the model's confidence is low; with the predicted probabilities being in the 0.3 to 0.5 range. A quick analysis with some basic audio augmentations reveals minimal impact, and we defer it to a future study with more comprehensive data to understand the reason behind this phenomenon.