Quantum-Infused Whisper: A Framework for Replacing Classical Components

Tapabrata Mondal[1], **Debjit Dhar**[1], **Soham Lahiri**[1], **Sivaji Bandyopadhyay**[1*]
[1]Jadavpur University, Kolkata, India
[*]**Correspondence:** sivaji.cse.ju@gmail.com

Abstract

We propose a compact hybrid quantum–classical extension of OpenAI's Whisper in which classical components are replaced by Quantum Convolutional Neural Networks (QCNN), Quantum LSTMs (QLSTM), and optional Quantum Adaptive Self-Attention (QASA). Log-mel spectrograms are angle-encoded and processed by QCNN kernels, whose outputs feed a Transformer encoder, while QLSTM-based decoding introduces quantum-enhanced temporal modeling. The design incorporates pretrained acoustic embeddings and is constrained to NISQ-feasible circuit depths and qubit counts. Although this work is primarily architectural, we provide a fully specified, reproducible evaluation plan using Speech Commands, LibriSpeech, and Common Voice, along with strong classical baselines and measurable hypotheses for assessing noise robustness, efficiency, and parameter sparsity. To our knowledge, this is the first hardware-aware, module-wise quantum replacement framework for Whisper.

1 Introduction

Quantum Natural Language Processing (QNLP) and Quantum Automatic Speech Recognition (QASR) explore how quantum information processing can enhance representation, inference, and learning for language and speech. Prior work suggests that quantum models may offer richer expressivity for structured linguistic tasks (Wiebe et al., 2019) and improved efficiency for operations that are expensive in classical deep learning. Early demonstrations, ranging from compositional distributional models compiled with toolkits such as `lambeq` (Kartsaklis et al., 2021) to QCNN-based speech pipelines have shown encouraging results but are typically limited to small datasets and shallow circuits due to NISQ constraints.

Current quantum hardware still imposes strict limits on circuit depth, qubit count, and data encoding, and full quantum replacements for attention, beam search, or large-scale sequence modeling remain largely unexplored. As a result, hybrid architectures that combine quantum modules with established classical components offer a practical interim path for advancing quantum-enhanced ASR.

In this work, we propose a unified quantum–augmented extension of Whisper in which classical convolution, recurrent, and attention blocks can be replaced with Quantum Convolutional Neural Networks (QCNN), Quantum LSTMs (QLSTM), and Quantum Adaptive Self-Attention (QASA). Our design is explicitly hardware-aware, specifying qubit requirements, depth-constrained variational layers, and angle-encoding strategies compatible with current NISQ devices. We further provide a rigorous and reproducible evaluation roadmap, including datasets, baselines, and measurable hypotheses, to quantify the potential benefits of quantum modules in robustness, sparsity, and low-resource performance. Finally, we outline the feasibility of implementing these components on present hardware through hybrid training, parameter-shift optimization, and noise-mitigation techniques. Our principal contributions are the following:

1. **Modular, integrable quantum replacements.** A hardware-aware framework that replaces Whisper's convolutional, recurrent and (optionally) attention blocks with QCNN, QLSTM and QASA modules — including concrete integration patterns (e.g., QLSTM gates inside a Transformer-style decoder with quantum outputs mapped to standard gating nonlinearities) and fallback hybrid strategies for QASA.
2. **NISQ-feasible specification + transfer-learning.** Per-module NISQ constraints (qubit budgets, circuit-depth limits, entangling

topologies, measurement channels, and angle/amplitude encoding) combined with pretrained acoustic embeddings so quantum layers refine high-level features have been provided.

2 Related Work

Quantum approaches to speech and language have expanded across recognition, classification, and generation. Miller et al. (Miller et al., 2024) fused STFT and LPC spectrograms, processing the LPC branch with a variational quantum circuit (VQC) before CNN-based classification, achieving 94.54% accuracy on Speech Commands (vs. 93.05% classical), with improved robustness and storage efficiency. Thejha et al. (Thejha et al., 2023) proposed a QCNN with CNOT gates and parameterized rotations (SX, SY, SZ) in Qiskit, reaching 99.10% accuracy for accent recognition (vs. 98.8% CNN). Wang et al. (Wang et al., 2023) combined WavLM-Large embeddings with a low-dimensional VQC for synthetic speech detection, improving equal error rate to 5.51% (vs. 6.80% baseline), highlighting the utility of quantum–embedding coupling.

In NLP, Yang et al. (Yang et al., 2022) introduced BERT-QTC, pairing a pretrained encoder with a quantum temporal convolution layer to enable federated learning privacy while improving intent classification accuracy (96.6% vs. 95.0%). Di Matteo et al. (Di Sipio et al., 2022) surveyed quantum-augmented NLP, showing QLSTMs and quantum Transformers achieve classical-level accuracy with fewer parameters, suggesting VQCs as efficient dense-layer replacements. Yang et al. (Yang et al., 2021) built a decentralized ASR pipeline where Mel-spectrograms pass through 2 × 2 QCNN kernels before a BiLSTM-attention model, reaching 95.12% accuracy with compact architectures.

Earlier, Fu et al. (Fu and Dai, 2009) integrated QNNs with particle-swarm optimization, reporting 84.5–85% accuracy on small-vocabulary tasks with faster, noise-resilient training. Pandey et al. (Pandey et al., 2023) introduced QLSTMs replacing gates with VQCs, outperforming classical LSTMs on code-mixed text but raising overfitting concerns. Abbaszade et al. (Abbaszade et al., 2023) applied DisCoCat-based quantum circuits to machine translation, achieving low error (MSE=0.0019) on English–Persian. Yoshimura et al. (Yoshimura et al., 2018) improved neural vocoders like WaveNet via mel-cepstrum quantization shaping, yielding a 0.6 MOS gain and 4 dB Equivalent-Q improvement with efficient MLSA filters.

Overall, these studies demonstrate the potential of hybrid quantum–classical methods for speech and NLP, spanning spectrogram fusion, quantum frontends, transfer learning, federated privacy, and model compression. In contrast, our work embeds parameterized quantum circuits directly into both feature extraction and decoding, integrates large pretrained acoustic embeddings for full transcription (not just classification/detection), and evaluates a cohesive end-to-end quantum–classical ASR pipeline on standard benchmarks, extending beyond earlier proof-of-concept systems.

3 Methodology

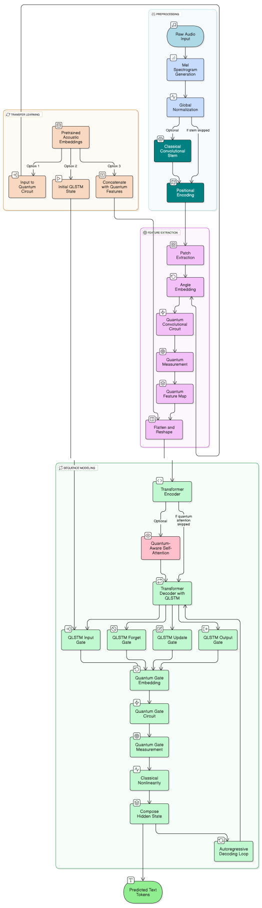

Figure 1: Architecture of the Quantum-Augmented Whisper pipeline. Log-mel patches are angle-encoded and processed by QCNN kernels that refine pretrained acoustic embeddings before a Transformer encoder; decoding uses QLSTM (VQCs replacing LSTM linear transforms) with optional QASA attention projections and a classical token head.

3.1 Feature Extraction with Quantum Convolutional Layers

An overview of the proposed Quantum-Augmented Whisper is shown in Figure 1, combining quantum and classical modules within the ASR pipeline. Raw audio at 16kHz is converted into an 80-channel log-Mel spectrogram using a 25ms window and 10ms stride, normalized and optionally processed by a lightweight convolutional stem with ReLU or GELU activations and positional encoding. For feature extraction, instead of classical CNNs we employ a Quantum Convolutional Neural Network (QCNN) (Yang et al., 2021), where 2×2 spectrogram patches are angle-encoded into 4-qubit states and processed by variational circuits with trainable rotations (R_X, R_Y, R_Z) and CNOT entanglement. Pauli-Z expectation values provide the quantum features, acting as trainable kernels that replace classical filters and are assembled into a quantum-enhanced feature map. This approach introduces stochasticity from measurement and exploits entanglement to capture local dependencies more effectively, particularly in low-data settings. While kernel sizes of 1×1 to 3×3 are considered, prior work indicates 2×2 offers the best trade-off. The resulting feature map is flattened into a temporal sequence for downstream modeling.

3.2 Whisper-Inspired Transformer Decoder with QLSTM Layers

Mathematical Foundation of Quantum LSTM Gates: Quantum Long Short-Term Memory (QLSTM) (Pandey et al., 2023) extends classical LSTMs by replacing linear transformations in gate computations with variational quantum circuits (VQCs). For each gate $g \in \{f, i, \tilde{C}, o\}$,

$$g_t^{(q)} = \sigma\big(\mathrm{VQC}_g\big([h_{t-1}, x_t]; \theta_g\big)\big), \tag{1}$$

where $[h_{t-1}, x_t]$ is the concatenated vector of previous hidden state and input, and θ_g are circuit parameters.

VQC Architecture: Each variational circuit operates in three stages:

1. *Encoding:* Inputs are mapped via angle encoding:

$$|\psi_{\text{enc}}\rangle = \bigotimes_{i=1}^{n} \Big(\cos \frac{\arctan v_{t,i}}{2} |0\rangle_i + \sin \frac{\arctan v_{t,i}}{2} |1\rangle_i\Big) \tag{2}$$

2. *Variational Layer:* For L layers and n qubits:

$$U_{\text{var}}(\theta) = \prod_{l=1}^{L} \left[\prod_{i=1}^{n} R(\alpha_{i,l}, \beta_{i,l}, \gamma_{i,l}) \prod_{\langle i,j \rangle} \mathrm{CNOT}_{i,j} \right], \tag{3}$$

where $R(\alpha, \beta, \gamma) = R_z(\gamma) R_y(\beta) R_z(\alpha)$.

3. *Measurement:* Expectation values are extracted on Pauli-Z:

$$\langle Z_i \rangle = \langle \psi_{\text{final}} | Z_i | \psi_{\text{final}} \rangle.$$

The resulting QLSTM dynamics are:

$$f_t = \sigma(\mathrm{VQC}_f([h_{t-1}, x_t]; \theta_f)), \tag{4}$$
$$i_t = \sigma(\mathrm{VQC}_i([h_{t-1}, x_t]; \theta_i)), \tag{5}$$
$$\tilde{C}_t = \tanh(\mathrm{VQC}_{\tilde{C}}([h_{t-1}, x_t]; \theta_{\tilde{C}})), \tag{6}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{7}$$
$$o_t = \sigma(\mathrm{VQC}_o([h_{t-1}, x_t]; \theta_o)), \tag{8}$$
$$h_t = \mathrm{VQC}_h(o_t \odot \tanh(C_t); \theta_h). \tag{9}$$

Integration into Whisper Decoder: As shown in Figure 1, quantum-enhanced acoustic embeddings are processed by an encoder–decoder Transformer modeled after Whisper. The encoder uses stacked self-attention and feedforward blocks, while the decoder integrates QLSTM layers interleaved with self-attention and cross-attention modules. Each gate is realized by a parameterized quantum circuit using entangling layers and rotation blocks, with nonlinear mappings (sigmoid/tanh) ensuring standard gating behavior. This hybrid architecture preserves LSTM temporal dynamics while embedding them in quantum feature spaces, enhancing robustness to overfitting and demonstrating competitive accuracy in low-resource and multilingual ASR settings.

3.3 Quantum-Aware Self-Attention Module

While the Whisper encoder–decoder backbone ensures strong sequence modeling, we explore *quantum-enhanced attention* via *Quantum Adaptive Self-Attention (QASA)*, where query–key interactions are processed through PQCs to generate attention weights. Alternatively, PQCs can modulate key, query, or value vectors, injecting noise-aware or entangled projections that complement QLSTM temporal modules.

Quantum Adaptive Self-Attention (QASA) replaces classical dot-product attention with parameterized quantum circuits operating on encoded queries and keys. Given input tokens $X \in \mathbb{R}^{T\times d}$:

$$h_i^{(q)} = \tanh(W_q h_i) \tag{10}$$

$$\text{QASA}(h_i^{(q)}) = h_i + W_o \cdot QC(h_i^{(q)} + t) \tag{11}$$

where t is temporal information and $QC(\cdot)$ is a parameterized quantum circuit.

Quantum Circuit Details:

- *Data Encoding:*

$$\forall i \in \{1, \dots, n\} : R_X(x_i), R_Z(x_i)$$

- *Variational Rotations:* Trainable per-layer $R_Y(\theta_{l,i}), R_Z(\phi_{l,i})$
- *Entanglement:* Circular CNOT topology: CNOT$(i \to (i+1) \bmod n)$
- *Measurement:*

$$QC(h^{(q)}) = [\langle Z_j \rangle]_{j=1}^{n}$$

Quantum Encoding in Attention:

Amplitude-Encoded Attention:

$$|\text{Attention}\rangle = \sum_{i,j} \alpha_{ij} |i\rangle \otimes |j\rangle$$

Angle-Encoded Attention:

$$R_Y(\theta_{ij})|0\rangle = \cos\left(\frac{\theta_{ij}}{2}\right)|0\rangle + \sin\left(\frac{\theta_{ij}}{2}\right)|1\rangle$$

where θ_{ij} encodes attention between tokens i and j.

Hybrid Encodings: Multi-resolution, adaptive, or hierarchical encoding strategies may be applied depending on the attention head or input characteristics.

This extended theoretical grounding and mathematical exposition provides a robust foundation for quantum sequential models and quantum self-attention within ASR, adhering to pure quantum NLP principles throughout.

3.4 Transfer Learning with Pretrained Acoustic Embeddings

To enhance generalization and reduce training costs, we integrate *pretrained acoustic embeddings*, following Whisper's large-scale training paradigm (Figure 1). Contextualized features from models such as wav2vec 2.0 or Whisper's encoder are fused with QCNN outputs to provide richer representations. These embeddings can be incorporated in three ways: (1) as direct inputs to the quantum circuit, (2) as initial QLSTM hidden states, or (3) concatenated with QCNN outputs before transformer encoding. Leveraging embeddings trained on large corpora provides a strong acoustic prior, allowing quantum layers to refine higher-level representations rather than relearn fundamental audio patterns; an especially beneficial strategy in NISQ-constrained settings.

4 Evaluation Plan and Conclusion

Although primarily architectural, this work delivers a concrete, reproducible implementation and evaluation roadmap. The system can be evaluated on the Quantum-Augmented Whisper pipeline on three ASR settings—keyword spotting (Speech Commands), large-vocabulary transcription (LibriSpeech) and multilingual recognition (Common Voice)—using identical per-module circuit constraints and simulators (Qiskit Aer, PennyLane-Lightning). Implementation highlights: angle-encode log-mel patches into QCNN kernels (4-qubit patch kernels, per-module budgets 8–16 qubits), map VQC measurement expectations to classical projections and gating nonlinearities (sigmoid/tanh) for QLSTM integration, and operate quantum layers on frozen pretrained acoustic embeddings so quantum circuits refine high-level features. Experiments follow a progressive instantiation path. Ideal simulator → noise-injected simulator → limited-shot NISQ runs with standard mitigation (measurement error mitigation, zero-noise extrapolation) and hybrid training (parameter-shift gradients / classical optimizers, minibatching, staged unfreezing). It is expected that our model will show noise robustness, sample efficiency, and effective parameter counts. Testable hypotheses. QCNN frontends will likely reduce CER by 1–3% in noisy conditions via entanglement-mediated feature mixing. QLSTM decoding will improve low-resource generalization and quantum modules will match competitive accuracy with fewer parameters.

Limitations

The proposed hybrid quantum–classical ASR architecture faces several limitations. Simulating QCNN, QLSTM, and QASA circuits is computationally expensive, while NISQ devices impose decoherence, gate errors, and strict depth limits not fully captured in simulation. Jointly optimizing pretrained acoustic embeddings with quantum layers remains challenging. To ensure feasibility, QCNN kernels are restricted to six entangling gates per patch, QLSTM layers use 12–14 variational parameters on 8–12 qubits, and QCNN operates on four qubits per patch keeping all modules within an 8–16 qubit budget compatible with current IBM and IonQ hardware. Training is assumed on simulators using parameter-shift rules, with noise-aware transpilation, measurement-error mitigation, and simple entanglement topologies to ensure NISQ compatibility. Full end-to-end deployment may still require circuit cutting or hybrid execution until larger, more reliable quantum processors become available. This work should therefore be viewed as a hardware-aware architectural framework, providing a roadmap for empirical validation as quantum technology evolves.

References

Mina Abbaszade, Mariam Zomorodi, Vahid Salari, and Philip Kurian. 2023. Toward quantum machine translation of syntactically distinct languages. *arXiv preprint arXiv:2307.16576*.

Riccardo Di Sipio, Jia-Hong Huang, Samuel Yen-Chi Chen, Stefano Mangini, and Marcel Worring. 2022. The dawn of quantum natural language processing. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8612–8616. IEEE.

Lihui Fu and Junfeng Dai. 2009. A speech recognition based on quantum neural networks trained by ipso. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 2, pages 477–481. IEEE.

Dimitri Kartsaklis, Ian Fan, Richie Yeung, Anna Pearson, Robin Lorenz, Alexis Toumi, Giovanni de Felice, Konstantinos Meichanetzidis, Stephen Clark, and Bob Coecke. 2021. lambeq: An efficient high-level python library for quantum nlp. *Preprint*, arXiv:2110.04236.

Leslie Miller, Tanay Kamlesh Patel, Glen Uehara, Salil Naik, and Andreas Spanias. 2024. Quantum machine learning and spectrogram fusion for speech recognition. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pages 674–678. IEEE.

Shyambabu Pandey, Nihar Jyoti Basisth, Tushar Sachan, Neha Kumari, and Partha Pakray. 2023. Quantum machine learning for natural language processing application. *Physica A: Statistical Mechanics and its Applications*, 627:129123.

B Thejha, S Yogeswari, A Vishalli, and J Jeyalakshmi. 2023. Speech recognition using quantum convolutional neural network. In *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 1–7. IEEE.

Ruoyu Wang, Jun Du, and Tian Gao. 2023. Quantum transfer learning using the large-scale unsupervised pre-trained model wavlm-large for synthetic speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nathan Wiebe, Alex Bocharov, Paul Smolensky, Matthias Troyer, and Krysta M Svore. 2019. Quantum language processing. *Preprint*, arXiv:1902.05162.

Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Pin-Yu Chen, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. 2021. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6523–6527. IEEE.

Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen. 2022. When bert meets quantum temporal convolution learning for text classification in heterogeneous computing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8602–8606. IEEE.

Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. 2018. Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1177–1184.