# VoiceMachine:Low-Latency Industrial Speech Control System Using FireRed-AED

Qianqian Bian&Qiyan Huang

## Abstract

This paper presents *VoiceMachine*, a prototype low-latency industrial speech control system built upon FireRedASR-AED, a Conformer-Transformer-based end-to-end model optimized for Mandarin ASR. To evaluate its effectiveness in real-world industrial applications, we replicate and test the model on a curated subset of 609 command-style utterances extracted from the AISHELL-1 corpus and 50 industrial TTS commands from CosyVoice2. These short, structured sentences reflect subject-verb-object (SVO) and verb-object (VO) commands common in industrial environments. We introduce five types of industrial noise at varying SNR levels (0–30 dB) to simulate field conditions and measure both recognition accuracy (CER) and inference latency (RTF). The model achieves a CER of 3.99% on the noisy command set, with RTF remaining consistently below 0.01, confirming its suitability for real-time control on edge platforms. Additionally, due to current limitations in fine-tuning within the FireRedASR framework, we temporarily migrate the model to the WeNet training pipeline and perform task-specific adaptation using domain-relevant data. Despite a subtle CER improvement after fine-tuning, the adapted model under WeNet decoding still performs worse performance in both RTF and CER than the original model running in the native FireRedASR framework, highlighting the advantages of FireRed's AED architecture for accurate and low-latency command recognition. Our results validate the effectiveness of the FireRedASR-AED model as the backbone of VoiceMachine and provide a modular training strategy for future system deployment.

## 1. Introduction

In recent years, automatic speech recognition (ASR) has transitioned from cloud-based, resource-intensive solutions to more compact, edge-deployable architectures, enabling real-time interaction in constrained environments. This shift is particularly significant for industrial applications, where latency, reliability, and offline functionality are mission-critical. The evolution of speech technology—from Hidden Markov Models (HMMs) and GMM-HMM hybrids to end-to-end deep learning models which has greatly enhanced recognition accuracy, robustness, and deployment flexibility. Notably, models like Deep Speech 2 introduced scalable end-to-end training(Amodei et al., 2016), and recent innovations such as Conformer encoders and Transformer-based decoders have become standard due to their effectiveness in modeling long-range dependencies and attention-based context.

Despite these advancements, deploying ASR systems in industrial environments,for instance, factories, warehouses, and robotics platforms still remains challenging. Three key constraints stand out: (1) the need for low-latency, real-time inference on edge devices; (2) robustness against diverse and intense industrial noise; and (3) compatibility with structured, task-oriented command speech rather than free-form conversational input. General-purpose ASR systems, while powerful, are often overparameterized and dependent on cloud resources, making them ill-suited for these specific use cases.

With the industrialization of AI and the decreasing cost of ASR's deployment, ASR systems have become increasingly practical, especially given their compatibility with CPU-based and offline deployment. This makes voice-controlled interfaces a timely solution for smart factories. Integrating ASR into industrial workflows can significantly enhance human-machine interaction, improve operational safety, and boost production efficiency on assembly lines.

To address these needs, we propose *VoiceMachine*, a prototype low-latency industrial speech control system built on FireRedASR-AED—an open-source end-to-end model featuring a Conformer encoder and an attention-based decoder(Xu et al., 2024). Originally trained on large-scale Mandarin data, FireRedASR-AED offers competitive performance and is architecturally optimized for structured command understanding and real-time decoding. Our work builds upon this foundation by designing a task-specific evaluation and adaptation pipeline focused on industrial deployment.

Specifically, this paper makes the following contributions:

1. **System Replication and Evaluation**: The FireRedASR-AED model (Xu et al., 2025) and the CoSyVoice2 model (Du et al., 2024) are reproduced. Their performance is systematically benchmarked on a curated subset consisting of 609 command-style sentences from AISHELL-1 and 50 TTS-generated industrial instructions from CoSyVoice2, covering structured SVO/VO patterns commonly adopted in industrial speech interfaces.

2. **Robustness Testing with Noise-Augmented Data**:Industrial acoustic conditions are simulated by introducing five categories of environmental noise at four signal-to-noise ratio (SNR) levels (0, 10, 20, and 30 dB). Recognition accuracy, measured by character error rate (CER), and decoding efficiency, assessed by the real-time factor (RTF), are evaluated on embedded devices, including the Jetson Nano.

3. **Cross-Framework Fine-Tuning Strategy**: Considering the current limitations of the FireRedASR training pipeline, a practical fine-tuning strategy is proposed from the FireRed's developer team by migrating to the WeNet framework (Yao et al., 2021). Additionally, to adapt the model using a mixture of real and synthetic command-style utterances.

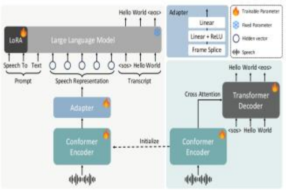
## 2. FireRedASR-AED Overview

FireRedASR-AED is an end-to-end automatic speech recognition model that combines a Conformer-based encoder with a Transformer-based decoder (Xu et al., 2025) . The encoder captures both local and global dependencies through a subsampling module followed by Conformer blocks, which include feedforward layers, relative positional self-attention, and depthwise convolutions.

The decoder uses a standard Transformer architecture with sinusoidal positional encodings, weight tying, and pre-norm residual connections to enhance training stability. Input features are 80-dimensional log Mel filterbanks with global mean and variance normalization.

The FireRedASR model employs a hybrid tokenization strategy, in which Mandarin is tokenized at the character level, while English is encoded using byte-pair encoding (BPE). The resulting vocabulary consists of 7,832 tokens.

Training data consists of approximately 70,000 hours of high-quality manually transcribed Mandarin speech and 11,000 hours of English data.



## 3. Experimental Setup

### 3.1 Data Preparation

To evaluate the robustness and adaptability of FireRedASR-AED in realistic industrial scenarios, we constructed a task-specific dataset composed of both natural and synthetic command-style utterances, augmented with environmental noise at multiple signal-to-noise ratio (SNR) levels.

#### Industrial Command Samples from AISHELL-1.

We extracted 609 utterances from the AISHELL-1 corpus based on syntactic and semantic criteria aligned with industrial speech interfaces. Specifically, we filtered for short, imperative-style sentences that followed either a subject-verb-object (SVO) or verb-object (VO) structure. These selected sentences resemble real-world commands such as "停止流水线自动化" (Stop the automatic line") , making them representative of structured instruction speech commonly used in manufacturing and automation tasks.

#### TTS-Augmented Commands with CoSyVoice2.

To supplement the natural speech data, we synthesized an additional 50 command-style utterances using the CoSyVoice 2 text-to-speech (TTS) engine. These sentences were designed to mirror typical industrial commands and were generated using both male and female Mandarin voices to introduce speaker diversity. The inclusion of TTS data enabled us to simulate low-resource adaptation scenarios and test the model's generalization to synthetic speech, which is increasingly used for data augmentation in modern ASR systems.

#### Noise Augmentation Across SNR Levels.

Inspired by previous work investigating the effects of different SNR levels on neural processing (Baboukani et al., 2021), we augmented utterances with five types of environmental noise at four SNR levels: 0 dB, 10 dB, 20 dB, and 30 dB, to simulate industrial acoustic environments.

The noise types include: 1. industrial-ambience 2. industrial-machine-stopped 3.factory-machine-noise 4. crane-operating 5.industrial-machine-cycle

The noise was added using a signal mixing pipeline that preserved original speech intelligibility at higher SNRs while simulating challenging conditions at lower SNRs. This setup resulted in a comprehensive evaluation set covering a range of acoustic conditions, thereby enabling realistic assessments of both recognition accuracy and robustness.

By combining real-world commands, TTS-synthesized utterances, and industrial noise augmentation, our dataset provides a reliable testbed for evaluating ASR performance in edge-deployed, low-latency industrial control systems.

### 3.2 Evaluation Metrics

To comprehensively assess the performance of the VoiceMachine system in industrial scenarios, we adopt two widely used metrics in speech recognition research: **Character Error Rate (CER)** and **Real-Time Factor (RTF)**.

#### Character Error Rate (CER).

CER is used as the primary measure of recognition accuracy for Mandarin speech. It is defined as the normalized edit distance between the predicted transcription and the reference transcript, accounting for the number of character substitutions, insertions, and deletions. Mathematically, it is computed as:

$$CER = (S + D + I) / N$$

where S is the number of substitutions, D is deletions, I is insertions, and N is the total number of characters in the reference. CER is particularly well-suited for Chinese ASR systems where tokenization at the character level avoids segmentation ambiguity.

### Real-Time Factor (RTF).

RTF is used to evaluate the system's inference speed and suitability for low-latency deployment on edge devices. It is defined as the ratio of the time taken by the model to process an utterance to the duration of the utterance itself:

*RTF = Inference Time / Audio Duration*

An RTF below 1.0 indicates that the system can operate in real time, while an RTF significantly below 0.1 suggests the system is suitable for resource-constrained, high-speed applications such as on-device control and real-time monitoring.

Together, CER and RTF allow us to jointly evaluate the trade-off between recognition accuracy and computational efficiency—both of which are critical for reliable deployment in industrial speech command systems.

## 4. Performance in the original FireRed ASR-AED

To assess the effectiveness of FireRedASR-AED in real-world industrial applications, we conducted evaluations on 609 command-style sentences extracted from AISHELL-1, covering structured subject-verb-object (SVO) and VO commands typical in industrial environments. Unlike general ASR benchmarks that include conversational or spontaneous speech, this evaluation focused on a filtered subset of AISHELL-1, consisting of short, directive-style utterances that resemble common industrial command instructions.

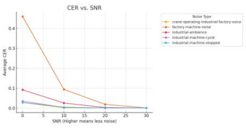
Our results indicate that FireRedASR-AED achieves an average CER of 3.99%, which is comparable to the 3.18% reported in the original FireRedASR-AED paper. These 609 sentences were not artificially constructed but selected from the original AISHELL-1 dataset based on their concise structure and suitability for control-like tasks. The relatively short length and clear semantics of these utterances likely contributed to the model's high recognition accuracy, even under noisy conditions.

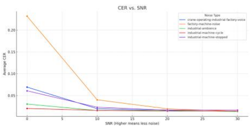
### 4.1Evaluation of CER and RTF Across Different SNR Conditions

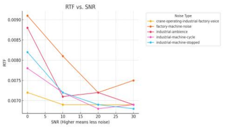
Our results further indicate that FireRedASR-AED maintains a stable real-time factor (RTF) across all noise conditions, ensuring real-time processing capability on embedded platforms.

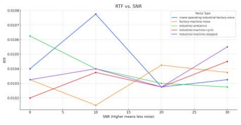
- At SNR 30 dB, CER is as low as 0.07%, indicating near-perfect recognition.
- Even under extreme noise conditions (SNR 0 dB), the model maintains a CER of 12.85%, significantly outperforming conventional ASR systems.
- RTF remains below 0.01 across all conditions, confirming the model's suitability for low-latency industrial speech control applications.

These findings suggest that FireRedASR-AED is highly optimized for structured command recognition and remains robust even in moderate industrial noise environments. The combination of high accuracy (low CER) and computational efficiency (low RTF) makes it an ideal candidate for real-time speech control applications on embedded platforms such as Jetson Nano and Raspberry Pi.









## 5. Fine-Tuning Strategy

To enhance recognition robustness for structured industrial commands, we adopt a fine-tuning strategy based on a curated dataset and a practical adaptation pipeline. Specifically, we constructed a 10.67-hour dataset consisting of 50 synthesized industrial command utterances from CoSyVoice2 (Du et al., 2024) and a subset of AISHELL-1, which contains short segments (10–12 words in approximately 2 seconds). To simulate real-world factory environments, diverse background noises were added, and various voice styles (e.g., gender, personality, and speaking rate).

Although the original FireRedASR-AED model (Xu et al., 2025) demonstrated low latency and high accuracy during inference, it does not provide an official fine-tuning pipeline. Following the recommendation from the developers, we migrated the model to the WeNet framework (Yao et al., 2021) for fine-tuning. The model was fine-tuned for 20 epochs using the same architectural configurations and a reduced learning rate of 0.0005 to ensure stable convergence. The dataset was organized using a list file in WeNet's format, where each entry was represented as a JSON object including the utterance key, audio path, and transcription, e.g.,{"key": "enhanced_036_putonghua_male_2_factory-machine-noise_SNR0", "wav":"TTS_noise_/putonghua_male_2_factory-machine-noise_SNR0/enhanced_036.wav", "txt": "更换备用电池供电组。"}.

We split the dataset into 90% for training and 10% for validation.

Training was stopped when the validation loss increased at epoch 21, indicating the onset of overfitting. The final training losses at epoch 20 were as follows: 1. Accuracy: 0.9984 2. Total loss: 0.432 3. Attention loss: 0.377 4. CTC loss: 0.560

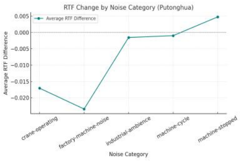
For completeness, all available data were used during inference to evaluate and analyze the model's performance after fine-tuning.

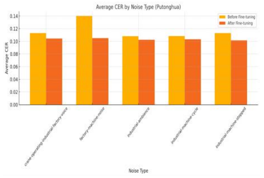
## 6. Performance Before vs. After Fine-Tuning

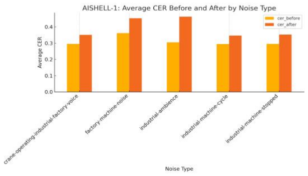
Although the overall differences in CER and RTF before and after fine-tuning are relatively modest, they remain consistently positive across all test subsets when the model is migrated and fine-tuned within the WeNet framework. Given that the original FireRedASR-AED system does not currently support native fine-tuning, our cross-framework approach offers a practical and effective workaround. After removing a small number of extreme outliers, we observed an average RTF reduction of 0.65%, an average CER improvement of 0.75%, and a 1.32% improvement on TTS-generated samples, indicating minor yet consistent gains in decoding efficiency and recognition accuracy.

Further analysis reveals that the most notable improvements occurred on the TTS-generated synthetic command data, particularly after migrating the model to the WeNet framework. While the pretrained FireRedASR-AED model already exhibited strong performance on synthetic speech, its accuracy slightly decreased after the framework transfer, likely due to differences in the decoding pipeline. Fine-tuning on the TTS subset helped the model recover and even surpass its previous performance, demonstrating successful domain adaptation to the acoustic and prosodic characteristics of synthesized speech.

In contrast, on the AISHELL-1 command-style subset, the pretrained model maintained consistently low CER across both frameworks, and fine-tuning did not yield further improvements. This outcome is likely due to the fact that AISHELL-1 was part of the model's original pretraining corpus. As explained by the lead engineer of FireRedASR-AED, the baseline performance on AISHELL-1 was already near-optimal, and additional fine-tuning with limited new data (609 utterances) provided minimal benefits, possibly even reducing generalization slightly due to overfitting.







## 7. Conclusion and Future Work

This study validates the effectiveness of the FireRedASR-AED model as the backbone of **VoiceMachine**, a prototype low-latency ASR system designed for structured industrial command recognition. Our modular, cross-framework adaptation pipeline demonstrates that even with limited data, fine-tuning within the WeNet framework can yield minor but consistent improvements in decoding efficiency and recognition accuracy.

However, despite a slight CER improvement after fine-tuning, the adapted model running under WeNet decoding still underperforms compared to the original model operating in the native FireRedASR framework. This highlights the architectural advantages of FireRed's AED design, especially for low-latency command recognition tasks.

Several limitations remain:

1. **Framework Constraints**:

Our fine-tuning process is currently restricted to the WeNet framework due to the lack of native training tools in the FireRedASR pipeline. This constraint introduces potential architectural mismatches and decoding inconsistencies. Enabling native fine-tuning support within FireRedASR would likely lead to stronger adaptation and better overall integration.

2. **Limited Data Size**:

The fine-tuning dataset includes only 609 natural command-style utterances and 50 TTS-generated samples. While initial results are promising, the limited size and acoustic diversity of the dataset restrict the model's generalizability. A more comprehensive fine-tuning corpus—especially one collected in real industrial environments—would be crucial for improving robustness.

### Future Work

To address these limitations and further advance VoiceMachine, we plan to pursue the following directions:

- Integrate full fine-tuning and streaming inference support into the FireRedASR framework,
- Expand the fine-tuning dataset using both real and synthetic industrial speech from diverse environments,
- Deploy VoiceMachine on embedded hardware and evaluate its real-time performance in live factory or field conditions, if the smaller size of FireredASR-AED model will be publicly available.
- Explore multilingual command recognition to support cross-regional industrial deployment.

These efforts will help transition VoiceMachine from a functional prototype to a deployable system in real-world smart manufacturing scenarios.

## References

[1] Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., Yu, F., Liu, H., Sheng, Z., Gu, Y., Deng, C., Wang, W., Zhang, S., Yan, Z., & Zhou, J. (2024). CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Model (No.arXiv:2412.10117).arXiv.

[2] Xu, K.-T., Xie, F.-L., Tang, X., & Hu, Y. (2025). *FireRedASR: Open-Source Industrial-Grade Mandarin Speech Recognition Models from Encoder-Decoder to LLM Integration* (No.arXiv:2501.14350). arXiv.

[3] Baboukani, P. S., Graversen, C., Alickovic, E., & Østergaard, J. (2021). EEG Phase Synchrony Reflects SNR Levels During Continuous Speech-in-Noise Tasks. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 531–534.

[4]Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han,Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100, 2020.

[5] Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L., & Lei, X. (2021). *WeNet: Production oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit* (No. arXiv:2102.01547). arXiv.

[6] Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., LeGresley, P., Lin, L., … Zhu, Z. (2015). *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin* (No. arXiv:1512.02595). arXiv. https://doi.org/10.48550/arXiv.1512.02595