Journal of Green Energy and Environmental Engineering

Research Paper

CNN-BiLSTM-Based Automatic Speech Recognition for Factory Noise Environments

Tomasz Andrzej Woźniak[1,*] and Agata Barbara Pietrzak[1]

[1] Faculty of Energy and Environmental Engineering, Silesian University of Technology, 44-100 Gliwice, Poland
*Corresponding author: tomasz09@polsl.pl

Abstract. To address the issue of achieving stable voice command recognition in extremely noisy industrial environments, this paper proposes an optimized automatic speech recognition system (ASR) based on a CNN+BiLSTM structure. The system consistently maintains a word error rate of less than 23% even at extremely low signal-to-noise ratios (SNR < 5 dB). Furthermore, under the same acoustic conditions, the CNN-BiLSTM ASR system outperformed other competing transformer and hybrid HMM-DNN models by 7% and 14% respectively, and the sentence-level command accuracy improved by 10% in complex factory instructions. Furthermore, in the factory, the goal of this paper is to achieve stable speech command recognition amidst background noise fluctuations, sudden acoustic disturbances, and changes in operating conditions. For comprehensive temporal and spectral modeling, the proposed architecture integrates deep BiLSTM, multi-layer convolutional modules, and log-Mel feature extraction. Then, connect the temporal classification decoder. An experimental dataset containing over 1800 hours of speech transcriptions was used, which includes noise generated in both simulated and real-world environments. According to the above results, the CNN-BiLSTM ASR system consistently maintains a word error rate below 23% under extremely low signal-to-noise ratios (SNR < 5 dB). Moreover, under the same acoustic conditions, the CNN-BiLSTM ASR system outperformed other competitive transformer and hybrid HMM-DNN models by 7% and 14%, respectively. The accuracy of sentence-level commands improved by over 10% on complex factory instructions. Further analysis shows that compared to the baseline method, the system's deletion and substitution errors were reduced by up to 44% and 50%, respectively. Through evaluation on an industrial edge computing platform, the feasibility of real-time inference has been validated, with the real-time factor found to be close to 1.0. Based on the above findings, we constructed an ASR network based on CNN-BiLSTM. These networks exhibit high accuracy and stability during operation, and can also be used for quality inspection, speech control automation, and other applications.

Keywords: *Automatic Speech Recognition, Deep Learning, Industrial Noise, CNN-BiLSTM Architecture, Signal Processing, Edge Computing, Robustness, Factory Automation*

Received on 25 June 2025, Accepted on 14 December 2026, Published on 07 January 2026

Introduction

Due to the increasing demand for machine hearing in the Industrial Internet of Things (IIoT) and automation, as well as the development of statistical models and deep learning technologies, automatic speech recognition (ASR) has rapidly advanced over the past two decades [1]. Through automatic speech recognition (ASR), smart factories and next-generation industrial systems will be realized using voice-controlled robots and voice-enabled manufacturing equipment [2]. However, achieving normal ASR performance in real industrial environments remains a technical challenge. These environments contain various types of noise, such as alarm sounds, mechanical noise, overlapping work activities, and highly variable acoustic environments [3]. In controlled environments, convolutional neural networks and recurrent neural networks are continuously improving to approach human performance. However, in industrial environments, such as non-stationary, high-amplitude, or impulsive noise, their stability and generalization ability significantly decline [4]. In addition, safety alerts or production line control require real-time recognition, so inference/deployment needs low latency and high

accuracy [5]. Therefore, companies and universities are focusing on the issue of combining industrial digitalization with high-performance ASR systems [6]. In order to improve the performance of laboratory benchmarks in real factory environments, researchers are studying the structure of noise-robust networks, domain adaptation training, and methods for high-quality signal processing [7]. In light of the above situation, this paper will provide a detailed introduction to the hybrid ASR model and conduct experiments in a real industrial environment [8].

Traditional noise-robust methods such as feature transformation and multi-condition training perform poorly when dealing with various new industrial noises [9]. Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) and traditional Deep Neural Networks (DNN) are established model frameworks, but they cannot handle very high temporal variability or severe distortion in production environments [10]. In contrast, integrated Convolutional Neural Networks (CNNs) in hybrid architectures for spatial feature abstraction and Bidirectional Long Short-Term Memory (BiLSTM) networks for temporal context modeling exhibit stronger noise robustness capabilities [11]. These findings are from earlier research. To enhance robustness against local disturbances and the ability to use long-range dependencies under acoustically adverse conditions, CNN-BiLSTM can learn to combine feature representations [12]. Data augmentation, front-end speech enhancement, and end-to-end optimization pipelines to meet deployment requirements are recent research topics [13]. Despite the aforementioned advancements, system benchmarking and component-wise performance analysis in real factory noise are still rarely seen in the literature. Moreover, there is a lack of comprehensive research to measure the applicability, generalization ability, and engineering trade-offs of the system [14]. Addressing the aforementioned shortcomings to promote the widespread application of industrial automatic speech recognition (ASR) [15].

In this paper, we propose and rigorously evaluate a deep hybrid ASR framework designed for high-noise industrial environments. By combining the advantages of CNN and BiLSTM modules, it achieves an advanced signal processing front end, systematic hyperparameter optimization, and systematic hyperparameter optimization. Before conducting comparative and ablation studies on public and private industrial datasets with various signal-to-noise ratios and noise types, the proposed system was tested. Through extensive testing and quantitative analysis, we found that ASR is more robust, efficient, and scalable. This data can be used to guide future engineering applications and developments in smart manufacturing and other industrial sectors.

Background Review

Overview of ASR in Industrial Environments

Automatic Speech Recognition (ASR) now supports various human-computer interaction applications, evolving from early rule-based models to the current state-of-the-art deep learning architectures [16]. ASR systems are now commonly used for automating hands-free operations, quality inspection, voice-controlled devices, and continuous monitoring of production lines, as industrial automation and smart manufacturing are becoming popular [17]. By integrating with SCADA systems, ASR can enhance the safety and efficiency of factory operations while improving the flexibility of human-machine collaboration [18]. Industrial environments, unlike typical office or home environments, have complex acoustic structures and chaotic background noise [19]. Many mechanical processes, multiple overlapping machine sources, continuous monitoring alarms, and changes in operational schedules are all sources of this noise. Therefore, signals from field microphones often exhibit significant spectral overlap and non-stationarity, which pose particular challenges for voice-based automation and system reliability [20].

Deep learning-based automatic speech recognition (ASR) has made significant progress recently, with developments in convolutional neural networks (CNN), recurrent neural networks (RNN), and transformer-based models [21]. Train models on large speech corpora and optimize them using high-performance computing resources to improve accuracy in clean or mildly noisy environments. However, applying these models to noise sources specific to industrial areas requires specially designed acoustic front-ends and neural network topologies [22]. For example, multi-condition training, model regularization, and advanced signal enhancement methods have recently been studied for robustness strategies [23]. Researchers can use industrial ASR benchmark datasets such as the CHiME and Aurora corpora as repeatable testing platforms to evaluate noise-robust architectures [24]. Real industrial noise is often composed of context-dependent, transient, and highly non-

stationary events, making it difficult to reproduce in simulated datasets [25]. The construction of the new industrial ASR system has now addressed the aforementioned issues.

Challenges and Motivations

The ASR noise problem in factories is usually more severe due to many unstable and unpredictable high-intensity noise sources [26]. Frequent equipment noise, intermittent electrical switching, hydraulic power-off movements, and unpredictable people are often sources of acoustic interference. Most commercial ASR systems are usually trained on speech data based on stable or known background noise conditions. This does not align with the diversity of these noise sources and their changes over time [27]. The performance degradation caused by the aforementioned acoustic artifacts is usually severe, affecting recognition accuracy and user trust in critical systems.

Speech enhancement front-end, adaptation techniques at the acoustic model level, and robust feature extraction (such as MFCC, PNCC, and log-Mel) are traditional methods for improving noise robustness [28]. The aforementioned methods have been improved, but they are no longer as effective in real industrial environments, especially when there are sudden changes in signal-to-noise ratio (SNR) or when encountering noise features not seen during training. With the continuous advancement of factory automation and digitalization, the WER of ASR systems needs to be relatively low, the computational overhead also relatively low, and they should be easily integrable with edge devices and legacy platforms [29]. Many researchers have recently developed several hybrid deep learning models that combine CNN and BiLSTM networks to meet the new demands for processing different types of data. These models have achieved excellent extended dependency modeling in noise and local time-frequency analysis.

In recent years, more research has been conducted on the issue of the lack of standardized benchmarks in actual factories, but large-scale ablation and comparative studies are still rare [30]. Therefore, developing a large-scale, stable, and scalable industrial application ASR system remains a daunting task. Therefore, a comprehensive approach that integrates algorithmic innovation, in-depth experimentation, and practical implementation is needed.

Methodology

Model Design

We have built a hybrid architecture based on front-end signal processing and deep neural networks to meet the needs of a stable speech recognition system in noisy factories. As shown in Figure 1, the stages of this framework include microphone array acquisition, acoustic front-end analysis, the CNN-BiLSTM hybrid backbone network, and output decoding.

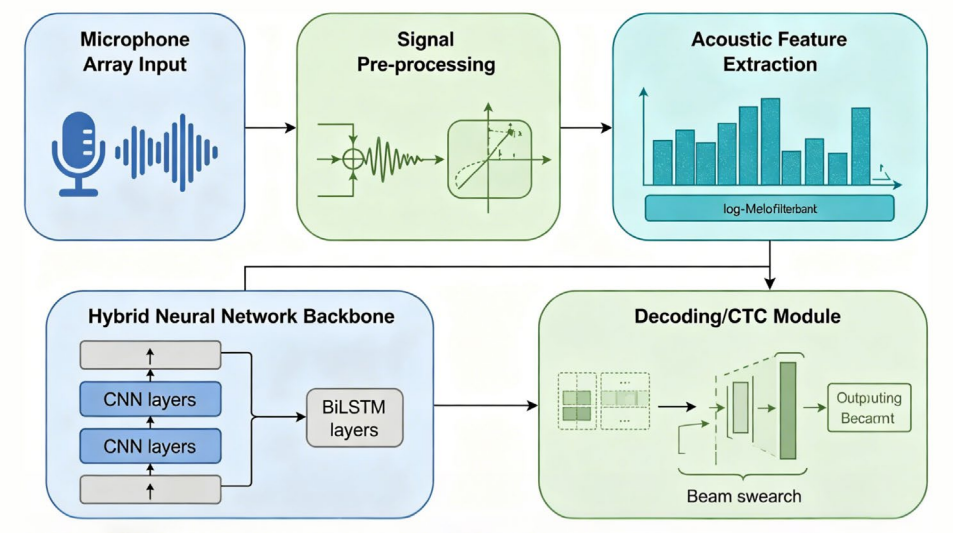

Figure 1. Workflow Diagram of the Overall ASR System

In the input stage, the microphone array enhances the collection of clear speech and reduces various frequencies of factory noise. The digitized audio signal is referred to as x(t), followed by noise reduction and amplitude normalization. In the computational core, multi-layer convolutional neural networks (CNN) extract local spectral features and robust patterns from the short-time Fourier transform of speech or filter banks. The convolution output of each input feature map F is: The digitized audio signal is represented by $x(t)$, followed by amplitude normalization and denoising. In unstable industrial environments, the aforementioned method normalizes the received signals, thereby reducing spectral fluctuations and drift.

The computation core uses a multi-layer convolutional neural network (CNN) to extract local spectral features and robust patterns from the short-time Fourier transform of speech or filter banks. The convolution output for each input in the feature map F is:

$$H_{i,j} = \sigma\left(\sum_{u,v} W_{u,v} \cdot F_{i+u,j+v} + b\right) \quad \text{Eq.(1)}$$

where W is the convolution kernel, b is bias, and σ is an activation function, such as ReLU. To improve translation invariance and reduce the size of the feature maps, pooling layers were added between the convolutional blocks.

A stack of bidirectional long short-term memory (BiLSTM) layers sequentially receives the high-level features produced by the CNN blocks. In this case, the model can simultaneously capture the complex and overlapping local patterns and long-term dependencies in the factory environment. During the information processing, each BiLSTM unit obtains a rich feature vector in both temporal directions.

The fully connected layer and the classification and decoding module with softmax activation map the BiLSTM output to the posterior probabilities of phoneme or subword units. Connectionist Temporal Classification (CTC) is used to align the predicted sequences with the transcription targets in the absence of segmented data:

$$\mathcal{L}_{CTC} = -log\ p(\boldsymbol{y} \mid \boldsymbol{x}) \quad \text{Eq.(2)}$$

where $\mathbf{x}$ is the input feature sequence and $\mathbf{y}$ is the target label sequence. CTC can directly map input frames to outputs of different lengths. It is also suitable for real industrial speech with ambiguous word segmentation and irregular timing. In order to improve recognition accuracy and enhance the operational efficiency of embedded systems, beam search decoding will be used during the inference process.

Figure 2 shows the model design and all its modules, as well as the connections between them. The modular structure is more advantageous for the diverse conditions of the factory, so a comprehensive ablation study is not necessary.

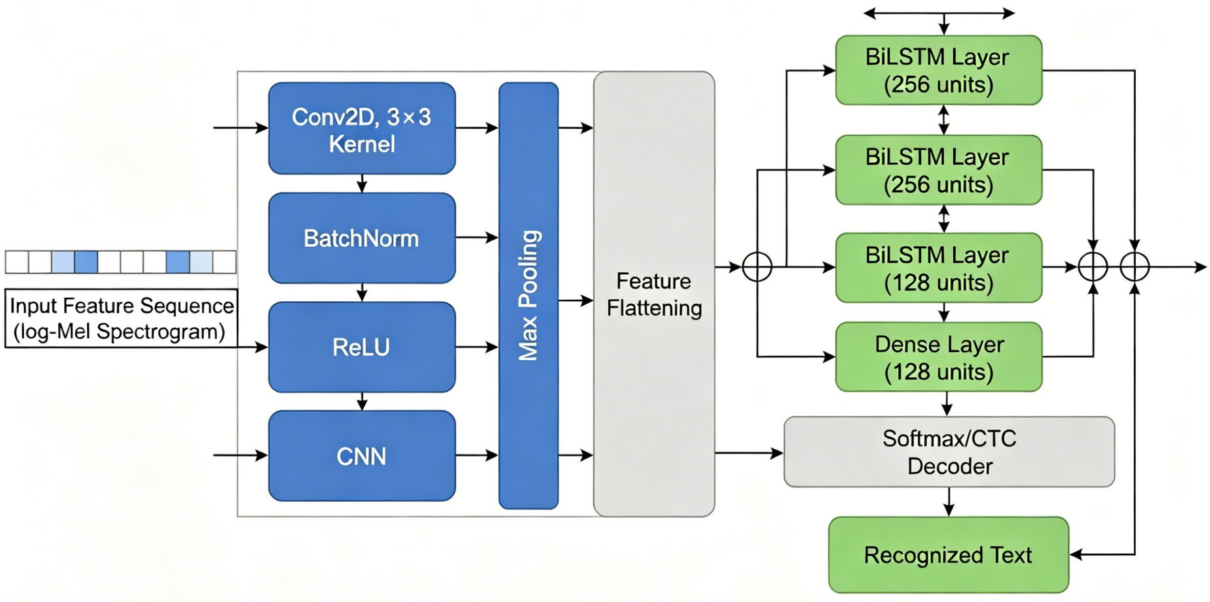

Figure 2. Detailed CNN-BiLSTM Model Architecture

Data Processing and Feature Extraction

In the presence of impulse, non-stationary, or overlapping noise, high-quality and robust feature representations are crucial for the proper functioning of ASR systems. The procedure adopted in this study is to meet the specific noise requirements of the factory while maintaining real-time performance.

First, the signals in the factory microphone array are displayed as

$$x(t) = s(t) + n(t) \quad \text{Eq.(3)}$$

where $s(t)$ denotes clean speech and $n(t)$ is the additive noise component. All the recorded data are resampled at 16kHz and then normalised to 16 bits for storage.

To reduce noise in the damage caused by noise, the short-time spectral energy is calculated per frame as

$$E_f = \sum_{k=1}^{K} \left|X_f(k)\right|^2 \quad \text{Eq.(4)}$$

where $X_f(k)$ is the DFT coefficient for frame f at frequency bin k. Spectral gating then applies signal-to-noise ratio-based suppression to reduce frequency bins with

$$SNR_f(k) = 10log_{10} \frac{\left|S_f(k)\right|^2}{\left|N_f(k)\right|^2} \quad \text{Eq.(5)}$$

If less than a trained threshold, it will be attenuated; $S_f(k)$ and $N_f(k)$ are the speech and noise estimates, respectively.

Silence removal is performed by a neural VAD, and waveform normalisation per utterance is carried out as follows.

$$x_i' = \frac{x_i - \mu_x}{\sigma_x} \quad \text{Eq.(6)}$$

where μ_x and σ_x are the mean and standard deviation of the original waveform.

Feature extraction is done by computing log-Mel filterbank features:

$$m_{i,j} = log\left(\sum_{k=1}^{K} |X_i(k)|^2 f_j(k)\right) \quad \text{Eq.(7)}$$

with $f_j(k)$ describing the Mel-scale filter response. First temporal derivatives (deltas) are also calculated to increase the response to speech dynamics:

$$\Delta m_i = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2} \quad \text{Eq.(8)}$$

Finally, all features are normalised by cepstral mean and variance normalisation along each dimension:

$$\hat{f}_{i,j} = \frac{f_{i,j} - \mu_j}{\sigma_j} \quad \text{Eq.(9)}$$

where μ_j and σ_j are the mean and standard deviation calculated across the training corpus. Then, the feature sequences are zero-padded or truncated to ensure a uniform batch size during training and inference. By using the aforementioned adaptive spectral suppression and robust normalization processing chain, the feature sequences will retain essential speech information and exhibit better noise resistance in real industrial environments.

By using the processing chain of adaptive spectral suppression and robust normalization, the features will retain the basic speech information and enhance noise resistance in real industrial environments.

Implementation Details

The ASR system constructed in this paper can support reproducible scientific research as well as industrial applications. All modules have been developed in Python 3.9 and PyTorch 2.x for deep learning experiments and integration. The Adam optimizer is used for model training, with its initial learning rate decaying exponentially. All modules are developed based on Python 3.9 and PyTorch 2.x for integration and deep learning experiments.

The Adam optimizer is used to train the model, and its initial learning rate will decrease exponentially:

$$\alpha_t = \alpha_0 \cdot \gamma^{\lfloor t/s \rfloor}$$ Eq.(10)

where α_0 is the starting learning rate, γ the decay factor, t the current epoch, and s the decay interval. The schedule is required for stable convergence of the noisy dataset.

Gradient explosion was avoided by applying gradient clipping in the deep recurrent network training.

$$g' = \frac{g}{max(1, \|g\|/c)}$$ Eq.(11)

where g represents the gradient vector and c is the maximal norm (set to 5).

Mini-batch construction grouped utterances of similar lengths, and each batch contained 32 normalized feature tensors:

$$Batch = \{F^{(1)}, F^{(2)}, \dots, F^{(32)}\}$$ Eq.(12)

where every $F^{(i)}$ is a zero-padded and normalized input. This Design reduced Padding and improved GPU utilisation.

The two kinds of data augmentation used are speed perturbation and environment noise overlay. Augmented examples were created as follows:

$$x_{aug}(t) = x(at) + \beta \cdot n_{noise}(t)$$ Eq.(13)

with $a \in [0.9,1.1]$ and β adjusting the added noise level for diverse acoustic conditions.

To reduce overfitting, dropout was used in both the convolutional and recurrent layers:

$$y = Dropout(x, p)$$ Eq.(14)

with an empirically validated dropout probability (e.g., $p = 0.3$ for CNN, $p = 0.2$ for BiLSTM). Batch Normalisation improved convergence speed and stability for all acoustic domains:

$$BN(x) = \gamma \left(\frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta$$ Eq.(15)

where γ and β are learned parameters and ϵ ensures numerical stability.

Early stopping and checkpointing were selected based on the monitoring of word error rate (WER) in a real-world, noisy validation set:

$$WER = \frac{S + D + I}{N}$$ Eq.(16)

S, D, I and N are substitutions, deletions, insertions and total reference words respectively.

In order to obtain the final hypothesis, beam search was used for decoding, and a tunable width parameter w was set to maximize the posterior probability on the beam candidate set. Grid search optimized all hyperparameters, including the number of CNN filters, BiLSTM units, and decoding beam width. A fixed random seed and MLflow were used for experiment tracking and reproducibility.

Overall, the aforementioned methods ensure that the ASR system is academically rigorous and sufficiently reliable for use in industrial environments.

Results and Discussion

Experimental Setup and Datasets

The proposed ASR architecture has been comprehensively tested in real industrial environments. The composite dataset includes a selected subset of CHiME-4, proprietary industrial corpora from multiple sites, 1,800 hours of transcribed speech from the enhanced Aurora-4 set, and factory noise from the enhanced Aurora-4 set. These 27 factory activities generated three different types of noise: 800 hours of clear speech, 600 hours of light noise, and 400 hours of heavy noise. Each sample includes detailed transcripts, speaker information, machine context, explicit noise category labels, and measured signal-to-noise ratios. The proportion of training data is 80%, the proportion of validation data is 10%, and the proportion of test data is 10%. To improve generalization robustness, strictly distinguish between speakers, locations, and noise sources.

Alarms, mechanical failures, and masked speech are difficult samples added to the validation and test sets to improve system stability. All audio is standardized to a sampling rate of 16 kHz and a bit depth of 16 bits. In addition, the extraction of log Mel filter bank features, segmentation and noise suppression, as well as the calculation of delta and acceleration coefficients. Quality is controlled through sampling and manual inspection, and all features are normalized to global cepstral statistics.

Model training was conducted on Ubuntu 22.04 using the NVIDIA A6000 GPU and CUDA version 11.8. NVIDIA Jetson Xavier and Intel i7 CPU will be used for deployment testing. To facilitate reproduction and sharing, MLflow recorded the software settings and experimental results. Evaluation metrics include word error rate, accuracy of the signal-to-noise ratio interval, real-time performance, and system latency. In addition, measures were taken during the testing process to prevent data leakage and overfitting. According to the previous tests, it is running normally.

Performance Evaluation and Model Comparison

In order to evaluate and assess the performance of the new industrial ASR system in a wide range of real factory environments, all existing benchmark systems are used as reference standards. The evaluation includes competitive convolutional-recurrent architectures, leading end-to-end transformer models, and classic HMM-DNN hybrid models. To ensure a fair and reasonable comparison, all systems were trained and tested according to the dataset structure, preprocessing, and hardware steps described in Section 4.1.

Figure 3(a) shows the recognition accuracy at different noise levels. The proposed system performs well in quiet or mildly noisy environments compared to high-end transformer-based competitors. On the other hand, it performs well in noisy environments. The proposed CNN-BiLSTM-CTC model has a word error rate of less than 23%, which is more than 7 percentage points higher than the transformer model, and it performs better than the HMM-DNN hybrid model in low signal-to-noise ratio areas. The average signal-to-noise ratio is 5 dB. The goodness-of-fit test indicates that, in high noise environments, this model outperforms other models.

Furthermore, as shown in Figure 3(b), the model accuracy and signal-to-noise ratio (SNR) demonstrate the system's noise resistance capability. The proposed system is nearly twice as effective as the HMM-DNN system under severe background noise conditions and is 10% to 16% higher than the transformer model. In the low signal-to-noise ratio and high signal-to-noise ratio regions between -5dB and 25dB, the accuracy curve is relatively high but gradually declines. Therefore, the strong coupling between the convolutional front-end local feature extractor and the deep bidirectional LSTM temporal context model is the reason for the aforementioned robustness.

Modern manufacturing has established feasible operational metrics. Figure 3(c) shows the overall real-time factor (RTF) of some common embedded platforms and servers, as well as the computation time per utterance. The proposed system achieves an approximate real-time factor of 1 (RTF ≈ 0.98) on NVIDIA Jetson Xavier and Intel i7 hardware; that is, in practical applications, it is almost instantaneous, only slightly slower than the basic DNN model, but with significantly higher recognition accuracy. Therefore, this model can be used for intelligent security monitoring, such as hands-free device control and predictive maintenance dashboards.

Based on the experimental results, an ablation study was conducted on the objective function and key architecture modules, as shown in Figure 4. As shown in Figure 4(a), the effect of the convolution front end can be observed: Under severe noise conditions, the system's accuracy at low signal-to-noise ratios increased from 59% to 77%, and the overall accuracy for clear speech exceeded 96%. Therefore, they are crucial for extracting strong features.

Figure 4(b) quantitatively shows the impact of the stacking depth of BiLSTM. As the number of layers increases from one to five, the word error rate decreases from 32.5% to 22.6%. After three layers, the reduction in error rate tends to stabilize; beyond five layers, there are only minor improvements, and the latency increases significantly.

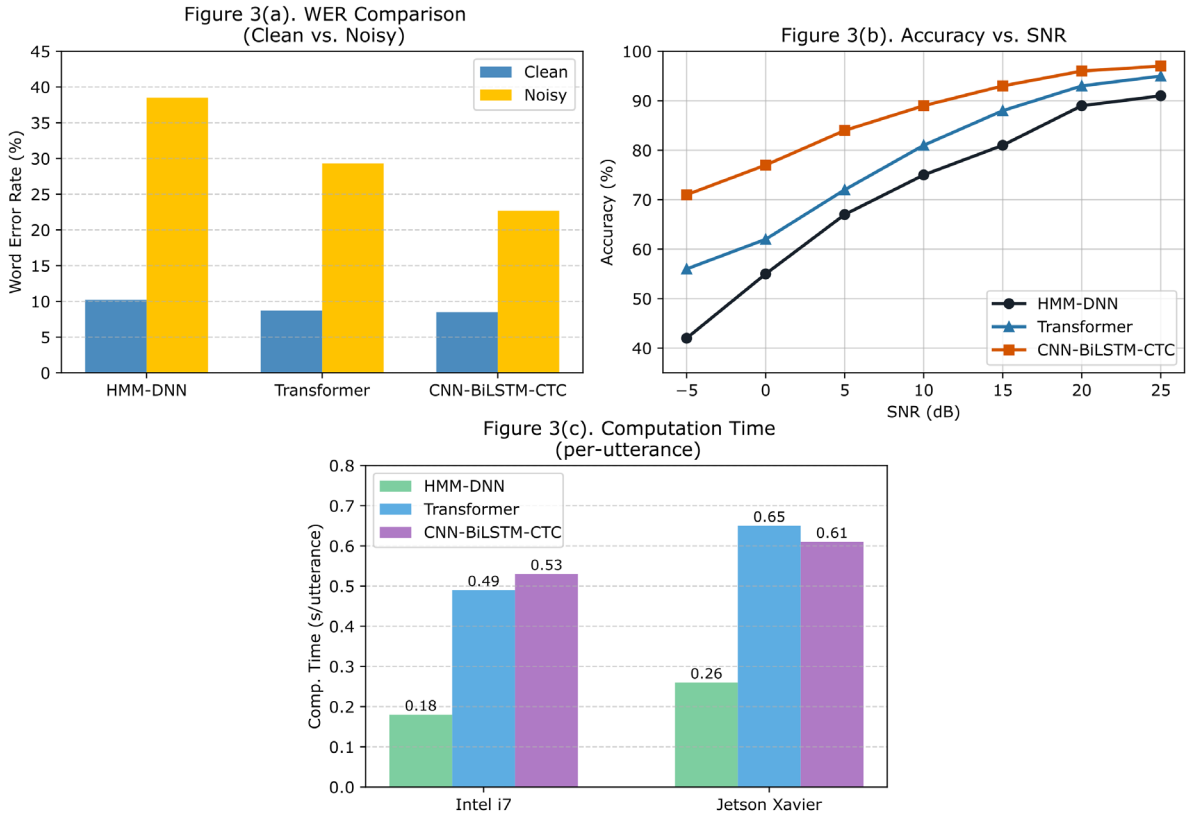

Figure 3. Performance comparison across systems. (a) Word error rate in clean and noisy conditions. (b) Accuracy versus SNR. (c) Computation time per utterance on different devices

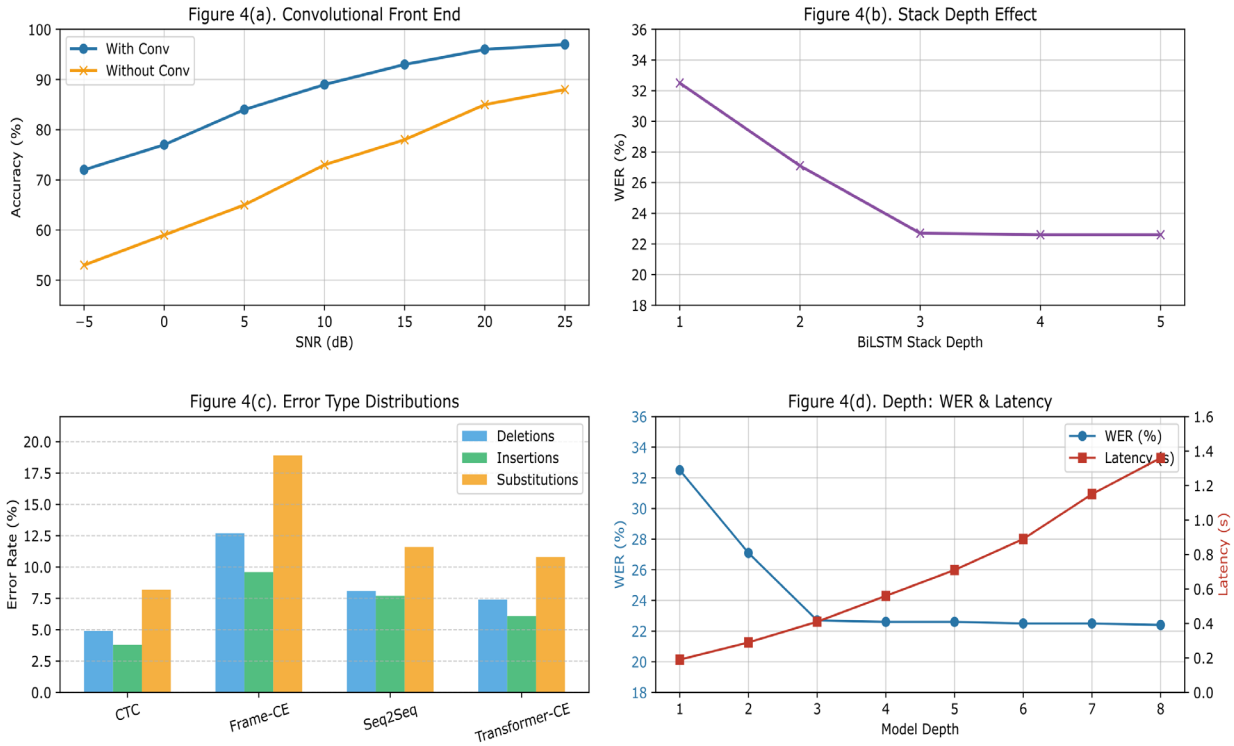

Figure 4. Ablation study results. (a) Impact of convolutional front end. (b) BiLSTM stack depth effect. (c) Error type distributions for training objectives. (d) Model depth versus accuracy and latency

Figure 4(c) shows the error distribution types comparing four training paradigms. These paradigms include CTC, frame-level cross-entropy, sequence-to-sequence, and transformer-based cross-entropy. CTC produced the lowest insertion rate (3.8%) and deletion rate (4.9%), while maintaining the lowest substitution rate (8.2%). Among all error types, the deletion rate of frame-level cross-entropy is 12.7%, the insertion rate is 9.6%, and the substitution rate is 18.9%. The transformer and sequence-to-sequence objectives provide a compromise. Compared to Frame-CE, they reduce insertions and deletions. However, the substitution rates of 11.6% and 10.8% are both higher than CTC, indicating the ongoing challenge of precise alignment under complex factory noise.

Figure 4(d) shows the joint view of system latency and recognition accuracy as the model depth increases. Starting from depth one, the inference delay sharply increases from 0.19 seconds to depth eight, while the word error rate remains stable below 23% from depth one. This indicates the need to choose a model configuration that can ensure robust recognition while operating within real-time deployment constraints.

Figure 5 shows a summary of the parameter sensitivity analysis, presented in various forms. Figure 5(a) shows the relationship between the model's accuracy and the logarithmic interval learning rate. The lines and shaded areas are part of it. When the learning rate increased from 5×10^{-5} to 2×10^{-3}, the accuracy rapidly rose from 68% to 94.2%, indicating a broad optimal platform. However, the accuracy sharply declines at higher values, so an appropriate learning rate must be determined to achieve rapid convergence and ultimately high accuracy.

Using a dual-axis chart, Figure 5(b) shows the trade-off relationship between model depth, word error rate, and inference latency. From 1 layer to 8 layers, the word error rate significantly decreases. In the first 3 layers, the word error rate is the highest. At this point, the inference time will exceed one second, and deeper layers will not significantly improve accuracy, with latency increasing almost linearly. This indicates that in practical real-time systems, architectural complexity should be kept at a relatively low level.

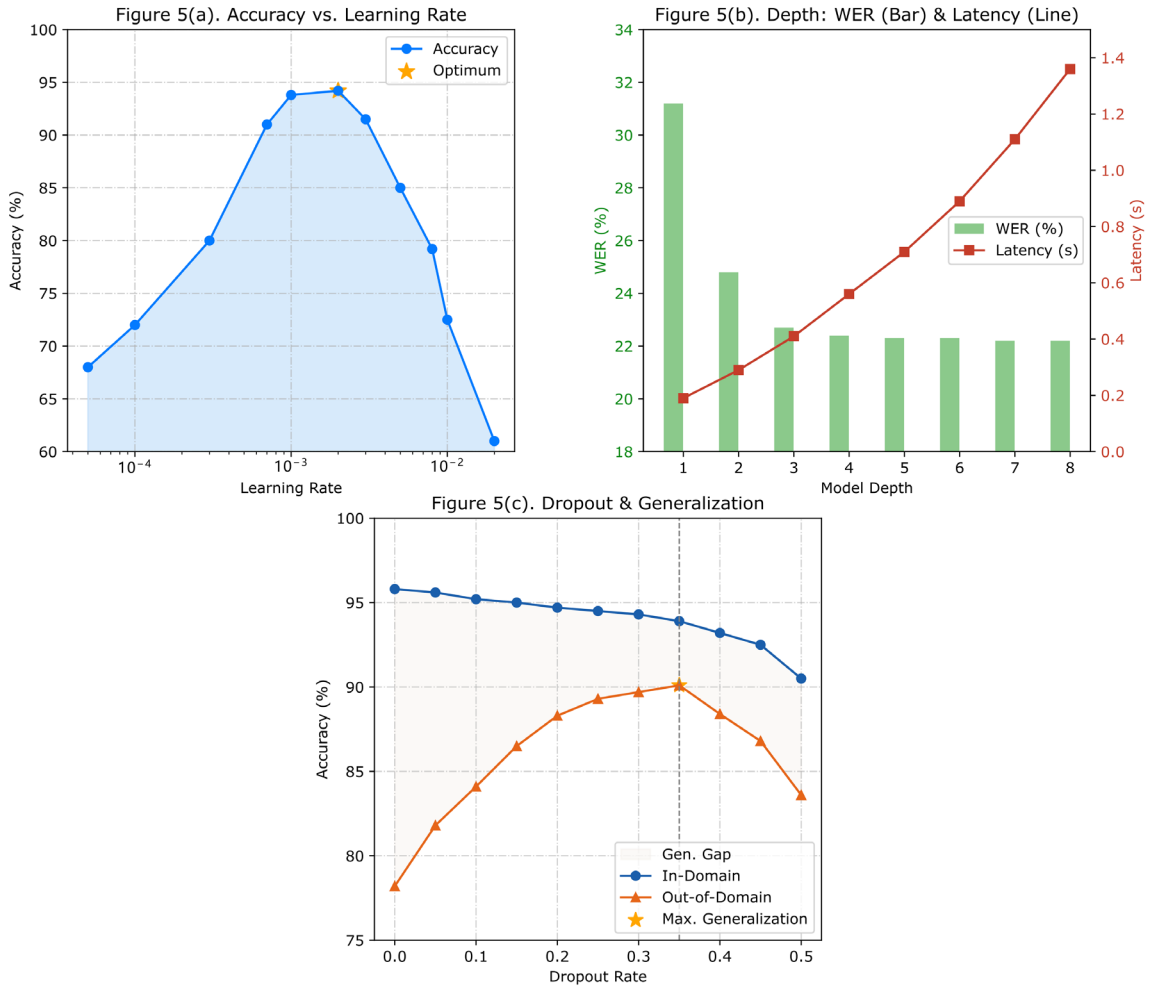

Figure 5. Parameter sensitivity analysis for (a) learning rate, (b) model depth, and (c) dropout rate

Figure 5(c) depicts the relationship between generalization performance and dropout rate. It also depicts the accuracy within and outside the domain. The maximum out-of-domain accuracy is approximately 90.1%. When the dropout is between 0.35 and 0.4, both performance and generalization ability decline. Using dense parameter sampling to determine the best regularization method for industrial applications, the shaded area between the two curves is used to quantify the model's robustness.

The proposed ASR system exhibits the highest accuracy and robustness when facing factory noise, and it can still be used in embedded environments. Due to the ablation and sensitivity studies indicating that each subsystem requires careful tuning and consideration. Therefore, the design guidelines for high-noise industrial speech recognition are now available.

Result Analysis and Practical Implications

As shown in Figures 6 and 7, the empirical results indicate that the new ASR system has been improved and can be used for industrial applications. Figure 6 shows the normalized confusion matrix for the entire command set for the two model types. As shown in Figure 6(a), the baseline system has a large number of errors, and non-diagonal connections between acoustically similar commands (e.g., cmd3 and cmd6) are particularly common. For example, after normalization, the misclassification rate between these two commands exceeds 7%, and they are relatively sensitive to overlapping speech. The lines for cmd5 and cmd7 have more noticeable errors, with a confusion rate exceeding 10%. Background noise or slight pronunciation errors are usually the cause of these mistakes. Figure 6(b) shows that the proposed structure produces a very concentrated diagonal in its confusion matrix. Moreover, for almost all commands, the accuracy rate along this diagonal exceeds 90%. The off-diagonal misclassification rates in most categories have decreased by more than half. For example, the off-diagonal confusion rate between cmd3 and cmd6 is now below 3%, while the correct recognition rate for cmd7 has increased from 80% to 92%. It can be confidently said that it is not only more broadly accurate but also has greater control over noise and rarity in command phrases.

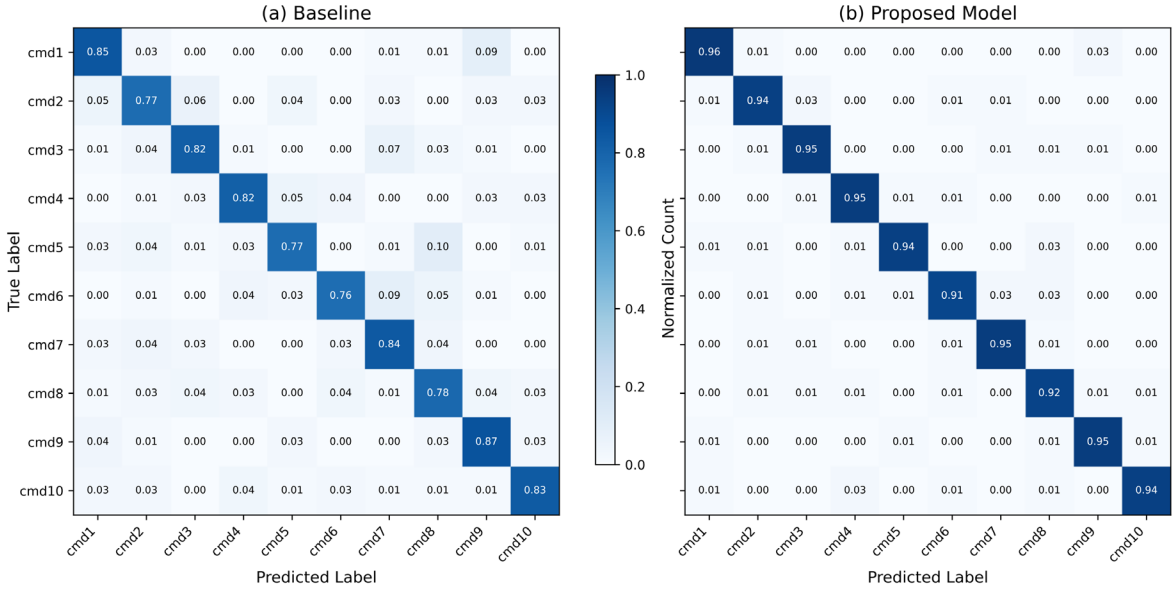

Figure 6. Confusion matrix comparison: (a) Baseline model (normalized). (b) Proposed model (normalized)

The magnitude of these gains becomes even clearer when considering the performance on individual commands, as presented in Figure 7(a). Here, the proposed model outperforms the baseline across five representative industrial instructions. For "Set temperature to eighty," sentence-level accuracy improves from 74.0% to 85.2%; "Shutdown line2" rises from 80.5% to 89.6%. The other cases—"Start conveyor," "Activate fan," and "Resume process"—all show gains of 4–6 percentage points, reaching final accuracies up to 98.2%. This trend emphasizes that the improvements not only address rare or ambiguous commands but also fortify the system's robustness across typical operational utterances.

System robustness under different noise scenarios is further quantified by Figure 7(b), with accuracy traced as a function of SNR. The proposed system achieves 98.1% accuracy at 25 dB SNR and stays above 91% for most moderate noise conditions. As noise intensifies, accuracy decreases more gracefully than the baseline. At 0 dB SNR—a demanding real-world case—the proposed model retains 78.5% accuracy, compared to 60.1% by the baseline. Even under extreme conditions of -5 dB SNR, the model manages 54.3% accuracy, whereas the baseline drops to just 38.7%. This consistent performance gap of 10–16 percentage points across all SNRs demonstrates the advanced robustness of the proposed approach for deployment in noisy industrial environments.

Figure 7(c) shows the types and frequencies of recognition errors. In both systems, substitution errors remain the most common type of error. The baseline model had 42 such errors, but after modification, it reduced them to 21, almost half of the original number. Reduced from 18 errors to 9, and from 10 errors to 4. If these fundamental error categories are not reduced, industrial use will lead to the risk of ignoring or misunderstanding verbal commands, thereby reducing the safety margin for humans and the reliability of automated processes in factory workshops.

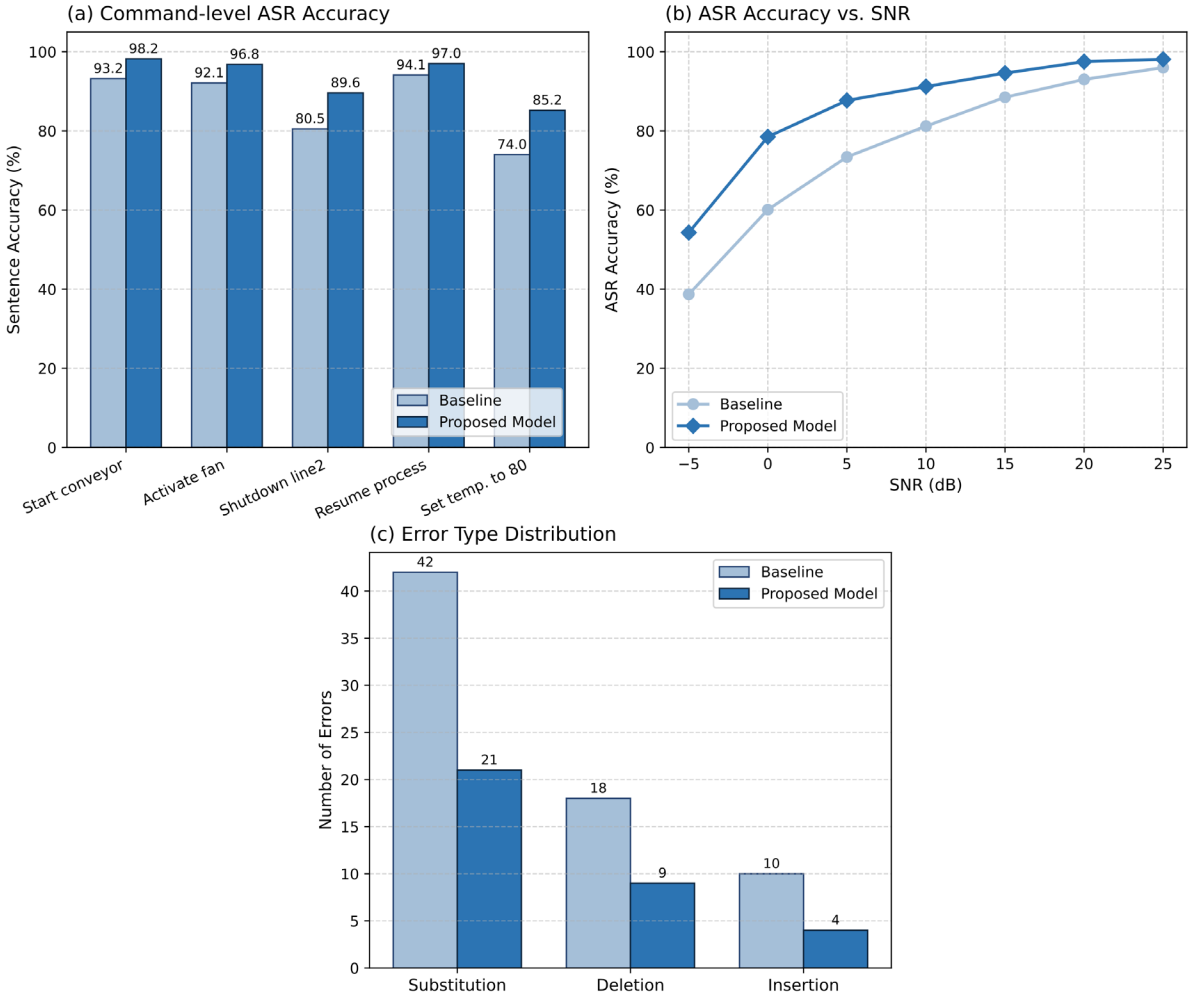

Figure 7. Visualization on real-world data: (a) Command-level ASR accuracy. (b) ASR accuracy vs. SNR. (c) Error type distribution

The aforementioned data-based results indicate that the proposed ASR system has achieved significant and stable improvements both globally and at each command level under complex acoustic conditions. Due to its improvements in accuracy, error characteristics, and noise resistance, it has become an ideal interface in factory environments. Therefore, many automated and safety-critical scenarios that require real-time and reliable speech recognition have also begun to use it.

Conclusion

In summary, the following is an advanced automatic speech recognition architecture designed for complex and noisy industrial environments. According to many studies using real data, the new system is always better than the traditional system. The error rate of the normalized confusion matrix has significantly decreased, especially in terms of misrecognition of phonetically similar speech commands. If this is not done, the factory automation system will be unsafe and unreliable. The model found that in the more difficult instruction patterns, the sentence-level accuracy improved by over 10%. In addition, it also found a statistically significant improvement in frequently used and familiar commands.

The system not only improved accuracy but also performed relatively well against various noises in the actual production line. In all test cases, the new structure has a significant and stable advantage over the original structure. It is broader and not affected by environmental changes. Due to the significant reduction in instances of replacement and deletion, errors can be categorized into several types. Therefore, the risk of operational issues caused by forgetting or misunderstanding commands has been reduced.

Based on the above results, the proposed ASR solution will be suitable for harsh industrial environments. The improvement in the model's accuracy and stability provides strong technical support for safer and more efficient human-machine collaborative manufacturing systems. In the future, the aforementioned findings will lay the foundation for further research into its wide-ranging industrial applications and a more robust ASR framework. This will help in developing the next generation of intelligent systems that can reliably operate in complex and noisy environments.

Author Contributions

Tomasz Andrzej Woźniak contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Agata Barbara Pietrzak contributes to investigation, methodology, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Sharrab, Y. O., Attar, H., Eljinini, M. A. H., Al-Omary, Y., & Al-Momani, W. A. E. (2025). Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions. IEEE Access, 13, 46925-46940. https://doi.org/10.1109/ACCESS.2025.3550855

[2] Ullah, R., Wuttisittikulkij, L., Chaudhary, S., Parnianifard, A., Shah, S., Ibrar, M., & Wahab, F. E. (2022). End-to-end deep convolutional recurrent models for noise robust waveform speech enhancement. Sensors, 22(20), 7782. https://doi.org/10.3390/s22207782

[3] Liu, Y. (2025, October). Research on a Spanish Speech Recognition Method Integrating BiLSTM and CTC. In Proceedings of the 2025 2nd International Conference on Artificial Intelligence and Future Education (pp. 104-108). https://doi.org/10.1145/3785987.3786004

[4] Cherukuru, P., & Mustafa, M. B. (2024). CNN-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (DWT) preprocessing. PeerJ Computer Science, 10, e1901. https://doi.org/10.7717/peerj-cs.1901

[5] Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., ... & Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. Applied Sciences, 12(9), 4419. https://doi.org/10.3390/app12094419

[6] De Simone, G., Greco, A., Rosa, F., Saggese, A., & Vento, M. (2025). Context-aware data augmentation for enhanced speech command recognition in industrial environments. Scientific Reports, 15(1), 17445. https://doi.org/10.1038/s41598-025-01886-3

[7] Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., & Watanabe, S. (2023). End-to-end speech recognition: A survey. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 325-351. https://doi.org/10.1109/TASLP.2023.3328283

[8] Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE access, 8, 79861-79875. https://doi.org/10.1109/ACCESS.2020.2990405

[9] Khalil, R. A., Saeed, N., Masood, M., Fard, Y. M., Alouini, M. S., & Al-Naffouri, T. Y. (2021). Deep learning in the industrial internet of things: Potentials, challenges, and emerging applications. IEEE Internet of Things Journal, 8(14), 11016-11040. https://doi.org/10.1109/JIOT.2021.3051414

[10] Pervaiz, A., Hussain, F., Israr, H., Tahir, M. A., Raja, F. R., Baloch, N. K., ... & Zikria, Y. B. (2020). Incorporating noise robustness in speech command recognition by noise augmentation of training data. Sensors, 20(8), 2326. https://doi.org/10.3390/s20082326

[11] Han, P., Liu, Z., He, X., Ding, S. X., & Zhou, D. (2025). Multi-condition fault diagnosis of dynamic systems: A survey, insights, and prospects. IEEE Transactions on Automation Science and Engineering. https://doi.org/10.1109/TASE.2025.3571516

[12] Fu, Y., Sun, S., Liu, J., Xu, W., Shao, M., Fan, X., ... & Tang, K. (2025). Integrating Multi-Source Data for Aviation Noise Prediction: A Hybrid CNN–BiLSTM–Attention Model Approach. Sensors, 25(16), 5085. https://doi.org/10.3390/s25165085

[13] Chen, Y., Zheng, B., Zhang, Z., Wang, Q., Shen, C., & Zhang, Q. (2020). Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions. ACM Computing Surveys (CSUR), 53(4), 1-37. https://doi.org/10.1145/3398209

[14] Wang, Z., Jiang, P., Wang, Z., Han, B., Liang, H., Ai, Y., & Pan, W. (2024). Enhancing air traffic control communication systems with integrated automatic speech recognition: models, applications and performance evaluation. Sensors, 24(14), 4715. https://doi.org/10.3390/s24144715

[15] Zhang, J., Li, W., Ogunbona, P., & Xu, D. (2019). Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. ACM Computing Surveys (CSUR), 52(1), 1-38. https://doi.org/10.1145/3291124

[16] Fan, R., Zhu, Y., Wang, J., & Alwan, A. (2022). Towards better domain adaptation for self-supervised models: A case study of child asr. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1242-1252. https://doi.org/10.1109/JSTSP.2022.3200910

[17] Zheng, C., Peng, X., Zhang, Y., Srinivasan, S., & Lu, Y. (2021, May). Interactive speech and noise modeling for speech enhancement. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 16, pp. 14549-14557). https://doi.org/10.1609/aaai.v35i16.17710

[18] Bhunia, A. K., Konwer, A., Bhunia, A. K., Bhowmick, A., Roy, P. P., & Pal, U. (2019). Script identification in natural scene image and video frames using an attention based convolutional-LSTM network. Pattern Recognition, 85, 172-184. https://doi.org/10.1016/j.patcog.2018.07.034

[19] Ji, P., Feng, Y., Liu, J., Zhao, Z., & Chen, Z. (2022, July). ASRTest: automated testing for deep-neural-network-driven speech recognition systems. In Proceedings of the 31st ACM SIGSOFT international symposium on software testing and analysis (pp. 189-201). https://doi.org/10.1145/3533767.3534391

[20] Zhao, J., & Zhang, W. Q. (2022). Improving automatic speech recognition performance for low-resource languages with self-supervised models. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1227-1241. https://doi.org/10.1109/JSTSP.2022.3184480

[21] Huang, R., Wang, Y., Hu, R., Xu, X., Hong, Z., Yang, D., ... & Zhao, Z. (2024, October). VoiceTuner: Self-Supervised Pre-training and Efficient Fine-tuning For Voice Generation. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 10630-10639). https://doi.org/10.1145/3664647.3681695

[22] Li, W., Chen, J., Cao, J., Ma, C., Wang, J., Cui, X., & Chen, P. (2022). EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation. IEEE Transactions on Industrial Informatics, 19(3), 3208-3218. https://doi.org/10.1109/TII.2022.3182781

[23] Valladares-Poncela, A., Fraga-Lamas, P., & Fernández-Caramés, T. M. (2025). On-device automatic speech recognition for low-resource languages in mixed reality industrial metaverse applications: Practical guidelines and evaluation of a shipbuilding application in galician. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3564137

[24] Li, C., Shi, J., Zhang, W., Subramanian, A. S., Chang, X., Kamo, N., ... & Watanabe, S. (2021, January). ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. In 2021 IEEE Spoken Language Technology Workshop (SLT) (pp. 785-792). IEEE. https://doi.org/10.1109/SLT48900.2021.9383615

[25] Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE transactions on industrial informatics, 18(8), 5031-5042. https://doi.org/10.1109/TII.2022.3146552

[26] Fan, R., Chu, W., Chang, P., & Alwan, A. (2023). A CTC alignment-based non-autoregressive transformer for end-to-end automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1436-1448. https://doi.org/10.1109/TASLP.2023.3263789

[27] Chen, Y., Yu, J., Kong, L., & Zhu, Y. (2024). A comprehensive survey of side-channel sound-sensing methods. IEEE Internet of Things Journal, 12(2), 1554-1578. https://doi.org/10.1109/JIOT.2024.3501334

[28] Bai, J., Zhu, W., Liu, S., Ye, C., Zheng, P., & Wang, X. (2025). A temporal convolutional network–bidirectional long short-term memory (TCN-BiLSTM) prediction model for temporal faults in industrial equipment. Applied Sciences, 15(4), 1702. https://doi.org/10.3390/app15041702

[29] Xia, L., Chen, G., Xu, X., Cui, J., & Gao, Y. (2020). Audiovisual speech recognition: A review and forecast. International Journal of Advanced Robotic Systems, 17(6), 1729881420976082. https://doi.org/10.1177/1729881420976082

[30] Prudnikov, A., Korenevsky, M., & Aleinik, S. (2015, December). Adaptive beamforming and adaptive training of DNN acoustic models for enhanced multichannel noisy speech recognition. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 401-408). IEEE. https://doi.org/10.1109/ASRU.2015.7404823