Article

Not peer-reviewed version

Optimizing the Human-Computer Interaction Interface of Warehouse Management Systems Using Automatic Speech Recognition Technology

Xue Song *

Posted Date: 15 November 2024

doi: 10.20944/preprints202411.1161.v1

Keywords: Speech Recognition; Adaptive Noise Suppression; Variational Autoencoder; Natural Language Processing

Article

Optimizing the Human-Computer Interaction Interface of Warehouse Management Systems Using Automatic Speech Recognition Technology

Xue Song

Department of Technology—User Experience, Fanatics Inc., San Mateo, CA 94403, USA; xsong@fanatics.com

Abstract: The exponential growth of the e-commerce and logistics industries in recent years has underscored the necessity for a warehouse management system (WMS) that is more efficient and intelligent. The prevailing WMSs are dependent on manual input and the use of handheld devices, which can result in inefficiencies and the potential for human error. In this work, we propose a human-computer interaction model for speech recognition optimized for the warehouse environment. This model integrates noise suppression technology based on an adaptive filter, which can dynamically detect and filter background noise, such as forklift operation sounds, human voices, and mechanical operation sounds. Furthermore, to address the variability in pronunciation, speech rate, and accent among different users, the system incorporates a speech enhancement model based on the variational autoencoder (VAE) technique. This approach enables the system to adaptively adjust the input speech features, thereby enhancing the robustness of recognition. In regard to natural language processing, a natural language understanding module based on bidirectional encoder representations from transformers (BERT) has been developed. The module is capable of semantic parsing of instructions in the user's voice and converting them into executable operation commands. Semantic slot filling technology enables the system to automatically identify the key entities in a task and perform linkage operations with the backend WMS database. The experimental analysis demonstrates that the proposed system is effective in an actual warehouse scenario. Compared with the traditional method, the task completion speed and accuracy are significantly improved.

Keywords: speech recognition; adaptive noise suppression; variational autoencoder; natural language processing

Introduction

The exponential growth of the e-commerce and logistics industries has led to the vital role that warehouse management systems (WMS) play in the modern supply chain. A WMS is primarily utilized to oversee and synchronize a multitude of operational procedures within a warehouse setting [1]. These encompass a range of activities, including inventory management, order picking, inbound and outbound logistics, and the formulation of logistics schedules. The efficiency of warehouse operations has a direct impact on the responsiveness and operating costs of the entire logistics system [2].

Consequently, enhancing the automation and intelligence of warehouse operations has become a pivotal strategy for modern enterprises seeking to enhance their competitiveness. However, conventional warehouse management techniques frequently depend on manual data entry and the use of handheld end devices. Despite its widespread use over the past few decades, this operational approach has become increasingly constrained by its inherent limitations. Manual operations are not only inefficient but also susceptible to error [3]. Operators must frequently switch devices and enter data manually, which not only increases operational time and costs but can also result in inventory errors and order delays due to human error. Furthermore, as the volume of orders rises, traditional handheld terminals are unable to fulfill the requirements for efficiency and accuracy.

In order to meet the growing business demands, it is imperative that enterprises implement a more efficient, convenient, and intelligent warehouse management solution. In recent years, the advancement of artificial intelligence (AI) and Internet of Things (IoT) technologies has led to the

 doi:10.20944/preprints202411.1161.v1

emergence of voice recognition technology as a prominent area of research within the field of warehouse management systems. Speech recognition technology allows warehouse operators to enter commands directly through their voice by translating speech signals into computer-interpretable instructions, obviating the necessity for a handheld device [4]. This feature provides the operator with a high degree of mobility, enabling them to perform a range of tasks, including item retrieval, inventory inquiry, order picking, and shelf replenishment, even when their hands are otherwise occupied. In this manner, the procedures associated with warehouse operations can be markedly optimized, which can enhance overall operational efficiency and mitigate the incidence of human error [5].

To illustrate, the conventional method of order picking necessitates that workers input confirmation data manually after locating an item on a shelf, a process that is inherently time-consuming. In contrast, voice recognition technology enables the operator to simply vocalize "Item A found," and the system records the completion. This not only reduces the time required for input but also minimizes the frequency of eye shifts between handheld devices and shelves, thereby enhancing the safety and efficiency of operations [6]. Furthermore, for tasks such as inventory inspection and shelf replenishment in warehouse management, voice recognition technology can assist operators in executing intricate operation instructions directly through voice.

Related Work

Initially, Zhang et al. [7] provide a comprehensive overview of recent advancements in robust speech recognition, with a particular focus on its application in noisy environments. They examine a range of deep learning techniques, including deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), which have been utilized to mitigate the adverse effects of noise on speech recognition accuracy. This is a particularly pertinent topic in warehouse settings, where background noise can pose a significant challenge.

In their study, Gajic et al. [8] address the issue of speech recognition in noisy settings. They introduce methods that leverage spectral centroid histograms to enhance the robustness of automatic speech recognition (ASR) systems. This approach is crucial for ensuring reliable operation in warehouses where machinery and ambient noise are prevalent.

Additionally, Qian et al. [9] investigate the potential of deep convolutional neural networks (DCNNs) to enhance speech recognition in challenging acoustic environments. They demonstrate how DCNNs can efficiently extract patterns and mitigate the impact of background noise, making them a promising approach for addressing the challenges posed by these environments.

The work of Kenton et al. [10] has had a profound impact on the field of natural language processing, particularly in terms of enhancing our ability to comprehend the contextual nuances embedded within sentences. This is of paramount importance in the context of warehouse management systems, which rely on accurate parsing of complex spoken commands to facilitate the execution of desired actions. The findings highlight the potential of pre-trained language models to be tailored for specific tasks, such as intent recognition and slot filling, which are crucial for the development of robust HCI systems.

Methodologies

Selecting a Template (Heading 2)

The adaptive noise reduction system is based on the adaptive filter of the least mean square (LMS) algorithm, which dynamically adjusts the filter coefficient to reduce noise. The noisy speech signal $x(t)$ received by the system is expressed as Equation 1.

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ represents the clean speech signal and $n(t)$ denotes the background noise. The objective is to estimate a clean speech signal $\hat{s}(t)$ by means of an adaptive filter $H(f)$, with the output expressed as per Equation 2.

 doi:10.20944/preprints202411.1161.v1

$$\hat{s}(t) = x(t) - \hat{n}(t), \tag{2}$$

where $\hat{n}(t)$ represents the estimated noise. The weight w of the adaptive filter is updated in accordance with the LMS update rule, as expressed by Equation 3.

$$w_{t+1} = w_t + 2\mu \cdot e(t) \cdot x(t), \tag{3}$$

where the filtering coefficient at time t is represented by w_t. The step size, or learning rate, is represented by μ. The error between the expected signal and the estimated signal is represented by $e(t)$, which is given by $e(t) = s(t) - \hat{s}(t)$. The input feature vector is represented by $x(t)$. The innovative aspect of this approach is the adaptive learning rate μ, which is adjusted in accordance with environmental conditions in order to achieve real-time adaptability, as illustrated by Equation 4.

$$\mu = \frac{\alpha}{\beta + \|x(t)\|^2}, \tag{4}$$

where the parameters α and β have been determined through experimentation with the objective of ensuring stability and rapid convergence. In an environment characterised by significant fluctuations in background noise, this adjustment enhances the efficacy of noise suppression.

The speech enhancement model addresses the issue of user pronunciation, accent and speech rate diversity through the use of a variational autoencoder. Given an input speech feature X that is subject to noise, the variational autoencoder learns a latent representation Z that enables the reconstruction of speech that is free from the effects of noise. The encoder maps the input X to a probabilistic latent space, as illustrated in Equation 5.

$$q_\emptyset(z|X) = \mathcal{N}\left(z; \mu_\emptyset(X), \sigma_\emptyset^2(X)\right), \tag{5}$$

where the parameter $\emptyset$ represents the encoder, while $\mu_\emptyset(X)$ and $\sigma_\emptyset^2(X)$ denote the mean and variance of the latent variables, respectively. The decoder is responsible for reconstructing the latent representation into speech features, as illustrated by Equation 6.

$$p_\theta(X|z) = \mathcal{N}\left(X; \hat{X}, \sigma^2\right), \tag{6}$$

In this context, θ represents the decoder parameter, while $\hat{X}$ denotes the reconstruction output. The objective of the optimisation process is to minimise the variational lower bound (ELBO), as illustrated in Equation 7.

$$\mathcal{L}(\theta, \emptyset; X) = -\mathbb{E}_{q_\emptyset(z|X)}[log p_\theta(X|z)] + D_{KL}\left(q_\emptyset(z|X) \parallel p(z)\right), \tag{7}$$

where D_{KL} represents the Kullback-Leibler divergence, and the distribution that encourages learning is in close proximity to the a priori $p(z)$. By means of dynamic prior adjustment, the system is able to make adaptive adjustments in accordance with alterations in the input speech mode, thus enhancing its adaptability to different pronunciations and speaking speeds and improving the robustness of the system.

Maintaining the Integrity of the Specifications

In the context of intent detection, the sentence representation $H = [h_1, h_2, \ldots, h_n]$, which has undergone BERT processing, employs the output h_{CLS} of the $[CLS]$ token as the representation vector for the entire sentence. The hcrs are then mapped to each intent class through a fully connected layer, and the probability is calculated by $softmax$, as illustrated in Equation 8.

$$p(I|x) = softmax(W_I h_{CLS} + b_I), \tag{8}$$

where the terms W_I and b_I refer to the weights and biases associated with intent detection. The final intent I is defined as the category with the highest probability, as illustrated by Equation 9.

$$I = argmax\ p(I|x). \tag{9}$$

 doi:10.20944/preprints202411.1161.v1

The slot filling task bears resemblance to sequence labelling, whereby a slot label is ascribed to each word of the input sentence. The BERT output for each word, designated as h, is subjected to further processing through the BiLSTM layer. This is done in order to capture long-distance dependencies, as illustrated by Equation 10.

$$h_i^{BiLSTM} = BiLSTM(h_i). \tag{10}$$

Subsequently, the fully connected layer is mapped to the labels of each slot, and the dependencies between the labels are ensured by the CRF (Conditional Random Field) layer, as illustrated in Equation 11.

$$p(S|x) = \frac{\prod_{i-1}^{n} \psi\left(s_{i-1}, s_i, h_i^{BiLSTM}\right)}{\sum_{s' \in S} \prod_{i-1}^{n} \psi\left(s'_{i-1}, s'_i, h_i^{BiLSTM}\right)}, \tag{11}$$

where $\psi\left(s_{i-1}, s_i, h_i^{BiLSTM}\right)$ represents a transfer function that calculates the transition probability between the two slot labels, s_{i-1} and s_i, in addition to the representation of the current position word. The CRF layer offers the advantage of integrating global sequence information, which enhances the coherence of bit prediction results. The loss function of the model is designed as a joint loss, with the objective of optimising both intent detection and bit filling. This is illustrated in Equation 12.

$$\mathcal{L} = \mathcal{L}_{intent} + \lambda \mathcal{L}_{solt}, \tag{12}$$

Note that the $\mathcal{L}_{intent}$ is obtained by following Equation 13.

$$\mathcal{L}_{intent} = -logp(I|x), \mathcal{L}_{solt} = -logp(S|x), \tag{13}$$

where the parameter designated as λ is responsible for regulating the combined weight of the aforementioned elements. By means of federated optimisation, the model is capable of disseminating pertinent information between intents and slots, thereby enhancing overall performance. The Slot Filling system is capable of automatically identifying the key entities within a given task, including product names, location numbers, and other pertinent information. The identified entities are linked to the backend WMS database in accordance with the following rules, as illustrated in Equation 14.

$$\begin{gathered} Q = SQL(I, S) = \\ SELECT\{S\} FROM\ WMS\ WHERE\ intent = I, \end{gathered} \tag{14}$$

where the variable Q represents the structured SQL query that has been generated based on the resolved intent I and the slot S.

Experimental

Experimental Setup

The experiment employs the Common Voice by Mozilla dataset, an open-source project spearheaded by Mozilla with the objective of amassing a diverse array of speech data. The database is now available in multiple languages and contains a substantial number of user-submitted voice samples with varying accents and speaking rates. The BERT model is configured as a 12-layer Transformer with 768 hidden cells per layer, utilizing an AdamW optimizer with learning rate decay, with an initial learning rate of $2e^{-5}$ and a batch size of 32. In order to prevent overfitting, a dropout probability of 0.1 is employed.

Experimental Analysis

Speech recognition accuracy represents a fundamental metric for evaluating the efficacy of speech recognition systems. Speech recognition accuracy is calculated as the ratio of the number of words that are incorrectly identified to the total number of words. The signal-to-noise ratio (SNR) represents a comparison of the recognition accuracy of a system under different background noise

 doi:10.20944/preprints202411.1161.v1

conditions. The signal-to-noise ratio is measured on a scale from 0 dB to 20 dB, with 0 dB indicating a noisy environment and 20 dB representing a relatively quiet environment.

Figure 1 provides a clear illustration of the performance of the "Ours" method under diverse SNR conditions, particularly at low SNR levels. However, the accuracy of alternative methods, including recurrent neural networks (RNNs), acoustic speech recognition (ASR), and deep convolutional neural networks (DCNNs), exhibited a notable decline at lower signal-to-noise ratios (SNRs). This demonstrates the robustness and advantage of our system in dealing with noise.

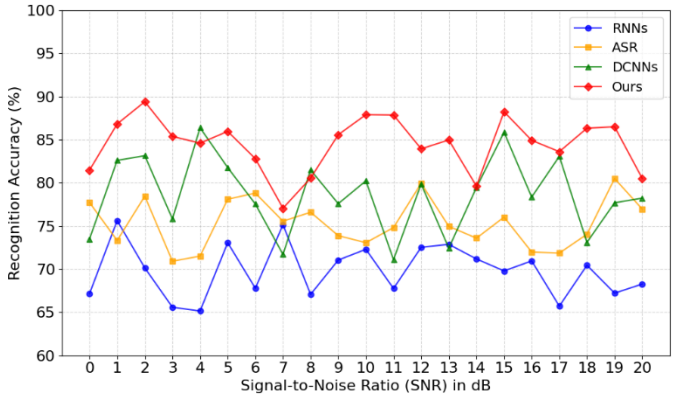

Figure 1. Recognition Accuracy Across Different SNR Levels.

Task Success Rate measures the ratio of the number of tasks successfully completed by the user through the system to the total number of tasks. This metric is a direct reflection of the reliability and effectiveness of the system in real-world applications, especially when handling critical tasks such as inventory queries, item positioning and order picking.

Figure 2 shows a comparison of task completion times for different methods under different signal-to-noise ratio (SNR) conditions. Task completion time: The time it takes for the user to give a voice command to the time it takes for the system to complete the task. As can be seen, the "Ours" method has short task completion times at all signal-to-noise ratios, especially at low SNR (noisy) conditions. In contrast, the task completion time of the other methods (RNNs, ASR, DCNNs) increases significantly with increasing noise levels, suggesting that our system has better robustness and responsiveness when dealing with noisy environments.

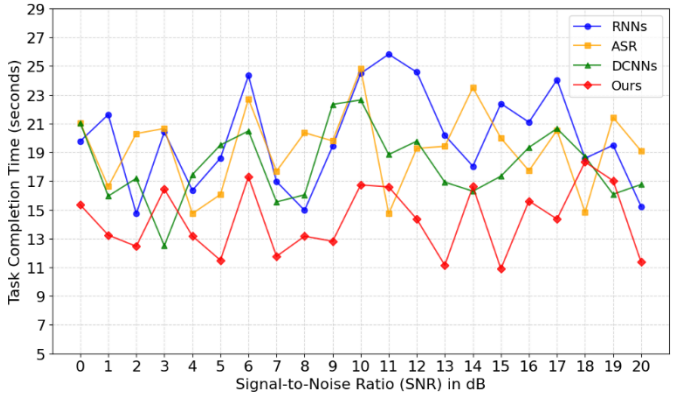

Figure 2. Task Completion Time Across Different SNR Levels.

The user satisfaction survey uses user feedback questionnaires to collect the subjective evaluation of the system experience of the warehouse operators participating in the trial, including ease of use, system response speed and detection accuracy. Satisfaction is rated on a five-point scale, ranging from 'very dissatisfied' to 'very satisfied', to provide a qualitative assessment of the overall user experience.

 doi:10.20944/preprints202411.1161.v1

Figure 3 shows a comparison of user satisfaction with different methods at different noise levels. The abscissa represents the level of noise intensity, from 1 (very low) to 5 (very high), and the ordinate represents user satisfaction, with a score ranging from 1 to 5, where 5 is very satisfied. As can be seen in Figure 3, our method ("our") maintained a high user satisfaction score at all noise intensities, even under high noise conditions. However, as the noise intensity increases, the user satisfaction of the other methods (RNNs, ASR, DCNNs) gradually decreases, especially under high noise conditions. This indicates that our system is more stable when dealing with background noise.

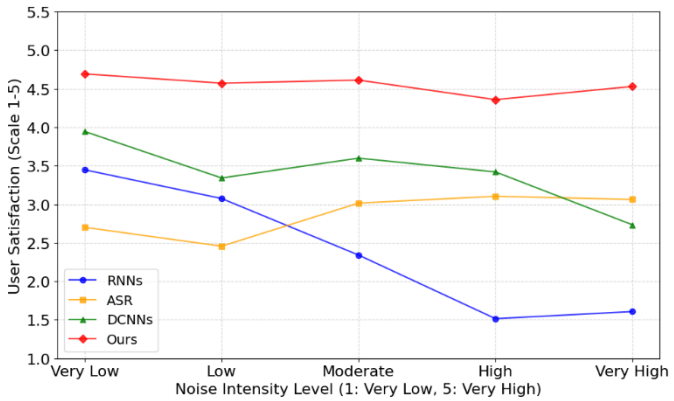

Figure 3. User Satisfaction Across Different Noise Levels.

Conclusions

In conclusion, we propose a human-computer interaction optimisation model for warehouse management systems, which is based on speech recognition technology with many innovations. Experimental results show that compared with the traditional RNN, ASR and DCNN methods, our method performs well in various experimental indicators. Whether it is the noisy warehouse environment, recognition accuracy, task completion efficiency or user satisfaction, it has shown higher robustness and practicality. Especially in the face of different background noise and diverse user input, our system is able to maintain stable high performance, significantly improving the efficiency and user experience of warehouse operations. Future research will optimise the real-time and cross-language compatibility of the system to ensure that it can operate stably in an international warehouse environment with multiple languages and accents. In addition, we will explore integration with vision technology to further improve the efficiency and accuracy of human-computer interaction through multimodal fusion.

References

1. Çimen, Egemen Berki, et al. "A Hybrid Stock optimization Approach for Inventory Management." 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2021.
2. Wang, Tingfei. "Construction and Data Integration of College Student Management Model Based on Human-Computer Interaction Data Acquisition and Monitoring System." Mobile Information Systems 2022.1 (2022): 9087983.
3. Döngül, Esra Sipahi, and Luigi Pio Leonardo Cavaliere. "Strategic management of platform business ecosystem using artificial intelligence supported human-computer interaction technology." Management and Information Technology in the Digital Era: Challenges and Perspectives. Emerald Publishing Limited, 2022. 47-61.
4. Shi, Jihua. "Research on Optimization of Cross-Border e-Commerce Logistics Distribution Network in the Context of Artificial Intelligence." Mobile Information Systems 2022.1 (2022): 3022280.
5. Manogaran, Gunasekaran, Chandu Thota, and Daphne Lopez. "Human-computer interaction with big data analytics." Research Anthology on Big Data Analytics, Architectures, and Applications. IGI global, 2022. 1578-1596.

6. Qi, Xiaoxuan, et al. "Intelligent retrieval method of power system service user satisfaction based on human-computer interaction." Journal of Interconnection Networks 22.Supp05 (2022): 2147012.
7. Zhang, Zixing, et al. "Deep learning for environmentally robust speech recognition: An overview of recent developments." ACM Transactions on Intelligent Systems and Technology (TIST) 9.5 (2018): 1-28.
8. Gajic, Bojana, and Kuldip K. Paliwal. "Robust speech recognition in noisy environments based on subband spectral centroid histograms." IEEE Transactions on Audio, Speech, and Language Processing 14.2 (2006): 600-608.
9. Qian, Yanmin, et al. "Very deep convolutional neural networks for noise robust speech recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.12 (2016): 2263-2276.
10. Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of naacL-HLT. Vol. 1. 2019.