Speech Interaction in Industrial Noise Scenarios

Yiyang Li

{ lyiyang@bupt.edu.cn }

International School Beijing University of Posts and Telecommunications, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. The application of speech recognition technology in the industrial field is becoming widespread. However, various occupational industrial background noises, characterized by high frequency and high loudness, have severely affected the performance of industrial speech interaction. Meanwhile, denoising technology has become a necessary step to popularize the application of speech recognition technology in industrial fields. This paper analyzes the basic attributes of different occupational noise in industrial scenarios. Focusing on the recognition and denoising modules in the industrial speech interaction process, it reviews the advantages and limitations of speech recognition technologies with traditional and deep learning-based baselines and denoising technologies based on speech enhancement. It proposed a collaborative industrial speech interaction model that includes perception, enhancement, recognition, and feedback layers. This paper aims to provide feasible strategies and further research guidance for the application of industrial speech interaction, offering significant practical values in this field.

Keywords: Human-computer interaction, Speech recognition in engineering applications, Speech enhancement, Noise suppression.

1 Introduction

The speech recognition technology has become a popular and widely used human-computer interaction technology and is applied to various scenarios in daily life including smart home [1], meeting content transcription [2], and real-time machine translation[2]. According to statistics, the global speech recognition software market grew from 7.268 billion US dollars in 2021 to 11.009 billion US dollars in 2025, and is expected to reach 25.262 billion US dollars by 2033, with a compound annual growth rate of 10.4% [3].

Nowadays, typical speech recognition technology has made remarkable research progress with a relatively high recognition accuracy rate. The speech recognition model developed by Song et al. based on the Hidden Markov Model (HMM) achieved a speech command recognition accuracy of up to 95% in a noise-free and quiet environment [4]. The recognition accuracy of the speech recognition model developed by Hamaza et al. using the Gaussian Mixture Model (GMM) reached 100% [5]. Radford et al. developed the Whisper model, which was trained on 680,000 hours of multilingual data and its speech recognition accuracy and robustness were

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365086

close to human-level performance [6]. In addition, the application of speech recognition technology is widely extending to various industrial scenarios and there are some differentiated demands. Rogowski emphasized the significant impact of background noise on the performance of speech recognition in industrial scenarios [7]. Compared to conversational patterns in daily life, communications in industrial work environments tend to be more concise. This requires that the speech recognition systems in industrial settings have a much lower tolerance for errors than those in everyday scenarios. According to Li et al.'s research, the interference on speech recognition caused by different occupational noises in industrial environments has scene specificity: Low-frequency piling noise in the mining industry (50-200Hz) can cause confusion of consonants between the command words "stop" and "start", while high-frequency concrete mixing noise in the construction industry (1000-3000Hz) can influence vowel recognition. Moreover, when the noise intensity exceeds 60dB, the accuracy of traditional convolutional neural network (CNN) speech recognition model drops by more than 30% [8].

Based on the aforementioned studies, the trend toward widespread adoption and daily use of speech recognition technology will inevitably lead to a continuous increase in the demand for this technology in industrial scenarios. Thus, this paper will systematically analyze the multi-occupational noise in industrial scenarios and its interference mechanism on speech recognition. It will focus on analyzing existing achievements in the speech enhancement denoising module and the speech recognition module during the industrial speech interaction process. Finally, it will propose a development concept for an integrated and collaborative optimization system for speech recognition in future industrial human-computer interaction scenarios.

This review aims to summarize and analyze the current research achievements of major speech denoising technologies and industrial speech recognition technologies, to propose a practical and effective technical guidance for industrial human-computer interaction, and to provide research pathways for the development of industrial speech interaction.

2 Basic attributes of industrial noise

2.1 Differences in the Sound Characteristics of Industrial Noise

Industrial environments consist of various occupational scenarios. According to NIOSH occupational hearing loss surveillance statistics, the average hearing loss rate is relatively high in four high-risk industrial fields: mining and oil/gas extraction, construction, agriculture/forestry, and manufacturing, which are all within the range of 20% to 30%[9].

Li et al. selected six types of industrial environmental noises as research objects, specifically including high and low-frequency piling, concrete mixer, hammering, weeder, and chainsaw [8]. All these noises are present in the four high-risk industrial scenarios mentioned above. They show significant differences in acoustic characteristics (e.g., spectral distribution, time-domain properties, and waveform harmonics). Their typicality and differences qualify them as typical cases for studying industrial noise. To verify their differences, this paper constructs a noise simulation model through Matlab to simulate and visualize the noise spectrum characteristics, as shown in Figure 1.

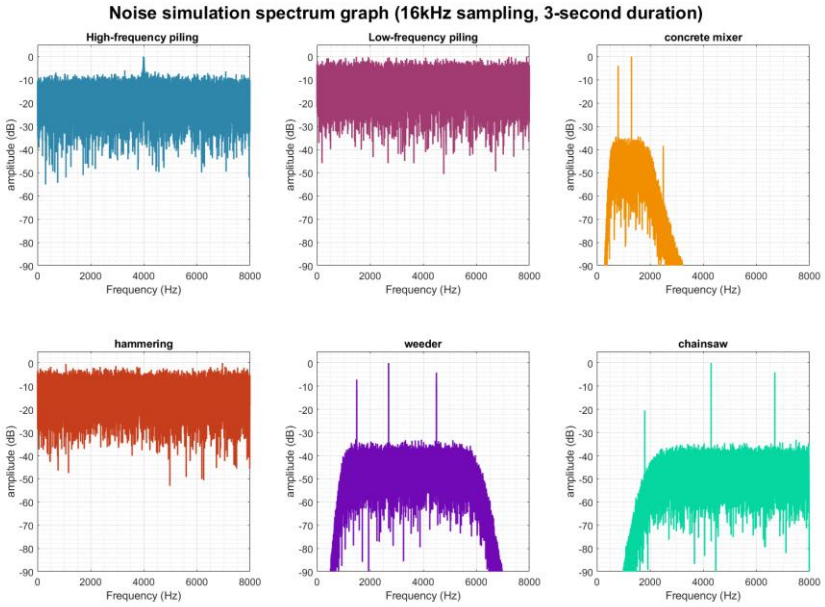

Fig. 1. Simulated frequency spectra of six typical industrial noises (Picture credit: Original)

2.2 Interference of Industrial Noise on Speech Recognition Models

Li et al. integrated the mentioned six types of typical noises as background noise into an industrial voice command speech data set and verified the noise interference under five noise intensity levels using a CNN speech recognition model. The results show that the accuracy of speech recognition in all background noise environments is lower than that in noise-free environments (89%). Moreover, significant differences in the accuracy among different types of industrial noises. For instance, in an environment of -20dB, the recognition accuracy difference between Hamming and weeder exceeds 10%[8]. It is evident that background noise in all industrial scenarios will reduce the accuracy of speech recognition models, and the extent of reduction is inversely correlated with the noise intensity level. Furthermore, ignoring the differences in noise intensity levels, the variations in noise categories can still lead to performance differences in speech recognition models, as illustrated in Figure 2. It can be inferred that the cause of this situation lies in the differences in core noise characteristics such as the frequency distribution and spectral properties. In summary, the differences in industrial noise types can lead to performance variations in speech recognition models, and the noise intensity levels can also affect their accuracy. Therefore, it is vital to distinguish different types of noise and denoise original voice signals in high-noise-level industrial environments to improve speech interaction performance in industrial scenarios.

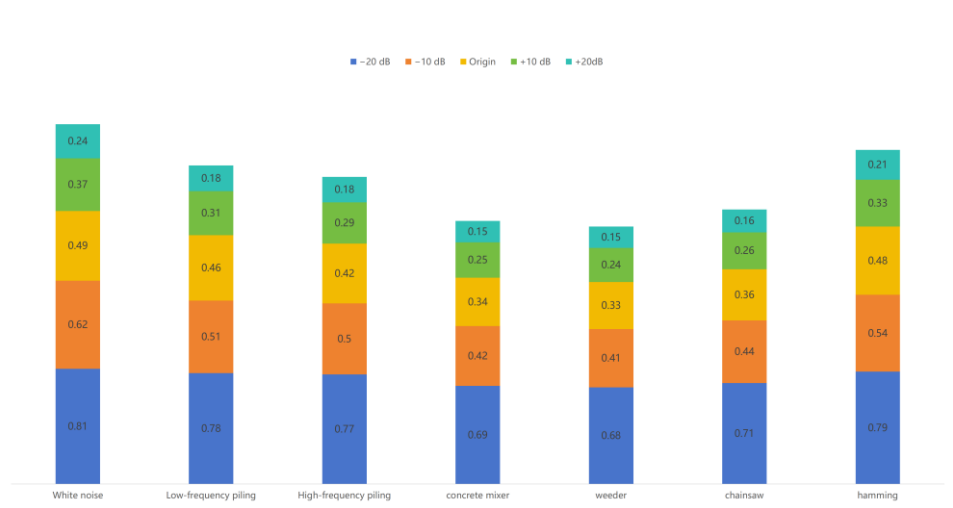

Fig. 2. Recognition accuracy based on the CNN model. (Data from [8])

3 Industrial Speech Interaction

3.1 Overview of the industrial speech interaction process

In industrial scenarios, the overall goal of speech interaction is to directly control the operation of industrial equipment through voice input. The specific process mainly includes: voice acquisition module: capturing the original voice signal through industrial sensitive microphone equipment. This original voice signal contains industrial speech instructions along with industrial environmental noise of high-intensity. This signal stream will be input into an industrial speech denoising module, which uses a specific speech denoising algorithm to reduce the noise and improve the SNR. This module will output denoised speech instructions, aiming to improve the accuracy of recognition in the following stage. The pure voice stream will then be input into the speech recognition module, where the speech instructions are analyzed and processed by specific algorithms, to match the preset speech instructions of the system and then output. Finally, the process enters the control response stage, where the interpreted command is sent to the control system, driving the device to perform the corresponding action, completing the conversion from speech to actual operation. Throughout the entire process, the result of the execution is fed back to the interaction system, establishing a closed-loop feedback mechanism with input, processing, and response stages.

3.2 Speech Denosing Module

Speech denoising, as a core module in industrial speech interaction, makes it possible to achieve reliable speech interaction in industrial environments with high-intensity background noise. Speech enhancement algorithms based on deep learning, by utilizing data-driven learning to distinguish noise from speech, perform exceptionally well in suppressing various types of industrial noises. They are currently the mainstream research direction in industrial noise reduction. The Deep Neural Networks (DNN) speech enhancement model proposed by

Xu et al. effectively improves the speech enhancement performance in dynamic noise scenarios, such as time-varying and impulsive noise [10], laying the foundation for the development of speech enhancement technology. The Speech Enhancement Generative Adversarial Network (SEGAN) developed by Pascua et al. is the first end-to-end wave-level speech enhancement model based on Generative Adversarial Network (GAN). This model adopts a fully convolutional encoder-decoder structure generator to convert the input noisy waveform into denoised speech; it uses discriminative network built by convolutional layers as a discriminator, which enables the generator to fine-tune the output signal waveform by conveying "real sample feature" information to be closer to the distribution of real speech. Ultimately, it achieves the denosing effect by identifying the noisy signal as the generated speech. In the tests involving 20 types of non-stationary noise, the SEGAN model achieved a segmental signal to noise ratio (SNR) of 7.73, showing its effectiveness in suppressing non-stationary noise in industrial scenarios [11]. Valin proposed RNNoise, a hybrid digital signal processor (DSP) / deep learning approach, which ensures real-time performance through DSP and improves noise estimation accuracy through recurrent neural network (RNN). The DSP decomposes the 16kHz full-band noisy speech into 20 overlapping sub-bands through a filter bank. At the same time, short-time energy normalization is performed on each sub-band to suppress the instantaneous interference of sudden large-amplitude noise. The lightweight RNN outputs the estimated value of the noise power spectrum for the sub-band, and leveraging the Wiener filter gain formula, see Equation (1).

$$G(k,t) = \max\left(1 - \frac{\hat{P}_N(k,t)}{\hat{P}_Y(k,t)}, \beta\right) \tag{1}$$

to calculate the noise suppression gain of the sub-band. Finally, the full-band speech is reconstructed through the sub-band synthesis filter to achieve denoising. RNNoise model owns lightweight features (0.8M parameters and less than 100KB of memory) and real-time response characteristics, which had an excellent denoising effect in NOISEX-92 dataset that the average signal noise ratio increase of 4.2 to 6.5dB [12].

Although single voice enhancement algorithm has indeed effectively addressed certain demands in some industrial speech interactions, there are still some problems. For instance, the SEGAN model achieves effective denoising, but its generative adversarial network inevitably has the problems of high latency and high deployment memory. Inversely, the RNNoise model is highly compatible with the requirements of embedded speech recognition devices in industrial scenarios, but it is relatively inferior in terms of enhancement effect [13].

3.3 Speech Recognition Module

The core of speech recognition technology is to transform continuous speech signals into correlated features that can be processed by machines, which requires algorithms to extract the features of speech signals. At present, mainstream speech recognition algorithms include two major technical pathways: traditional methods and deep learning-based methods, both of which have their own adaptability in industrial speech interaction scenarios.

Traditional Speech Recognition Algorithms. Traditional speech recognition algorithms rely on manual feature extraction and back-end modeling, which achieves speech recognition by jointly modeling the extracted acoustic features with GMM and HMM. Birch et al. adopted a

two-stage architecture of feature extraction and time series matching. The input audio undergoes "speech segment detection". The segments above the threshold are transformed into frequency spectrum using Fast Fourier Transform (FFT), and 13 correlated coefficients are then extracted as features through Mel-frequency Cepstral Coefficients (MFCC). Subsequently, the Dynamic Time Warping (DTW) algorithm computes the distance between the above coefficients and the coefficients in the sample dictionary to infer the target speech command. Integrated testing via the KUKA KR16 robot's RSI interface for this model has confirmed its stability in low-noise environments, achieving an accuracy rate of 93.51% under the condition that industrial environmental noise $\leqslant$ 61.2dB [14]. Long et al. optimized the traditional cochlear filter cepstral coefficient (CFCC) in two dimensions by replacing the conventional Fourier transform with fractional wavelet transform and logarithmic compression with non-linear power function compression. Then, they removed interference offsets through cepstrum mean normalization (CMN) and constructed high-dimensional feature vectors. These feature vectors were then processed by a support vector machine (SVM) classifier and combined with a dedicated acoustic model to ultimately achieve language recognition. Compared with the traditional CFCC algorithm, the false recognition rate of this algorithm under the mechanical noise conditions in the workshop has been reduced by 15.7%. Its performance has been significantly improved. Its performance has been significantly improved [15].

The advantages of traditional speech recognition algorithms mainly include low computing power requirements, short real-time latency, compatibility with industrial embedded device terminals, and low deployment costs. However, this type of model performs poorly in complex and high-noise environments. Thus, it is only applicable to industrial speech interaction scenarios where the noise is stable with relatively low intensity.

Deep learning-based speech recognition algorithms. Speech recognition algorithms based on deep learning use automatic feature learning methods and are data-driven. By utilizing neural networks, they can directly map original voice signals to text without requiring manual feature design. Hence, they achieve end-to-end speech recognition. Amodei et al.proposed the DeepSpeech2 speech recognition model, which adopts an end-to-end model architecture consisting of input feature processing, CNN, bidirectional long short-term memory network (BLSTM), fully connected layer and connectionist temporal classification (CTC) decoding. In the LibriSpeech general speech dataset test, the word error rate (WER) was as low as 6.9% [13]. This model achieves a breakthrough without the demand for manual feature extraction engineering, providing a high-quality research pathway for the industrial deployment and adaptive optimization of end-to-end learning models. Li et al. developed the Kaldi-ASR deep learning speech recognition model by integrating the deepSpeech2 with a front-end denoising module. Compared with the classic CNN speech recognition model, it had an average improvement in recognition accuracy from 14% to 35% [8]. This illustrates that the deep learning speech recognition models have excellent targeted optimization and deployment capabilities in industrial voice interaction scenarios.

The advantages of deep learning speech recognition algorithm models lie in their ability to achieve direct end-to-end mapping and can be specifically optimized for different industrial scenarios which have a better recognition effect in general. However, they also have limitations such as high computing power demand and relatively high latency.

4 Discussion

In industrial scenarios, both the speech denoising module and the speech recognition module should be adaptively optimized based on specific environmental noise. Therefore, as a pre-decision stage for speech recognition and denoising algorithms, noise recognition needs to determine the noise type accurately to provide a foundation for subsequent adaptive optimization. Based on the multi-occupational noise dimensions in industrial scenarios, this paper suggests that future research on noise recognition technology can focus on scene-noise correlation modeling, which is to construct a "scene-noise" mapping dictionaryto assist in identifying the dominant noise type and its characteristics through scene information.

Speech enhancement technology serves as an effective denoising approach in high-intensity background noise environments. The existing speech enhancement technologies target two core demands, lightweight (RNNoise) and high precision (SERAN), as the research goals and have achieved notable research results in both directions [10,11]. However, the complexity and high intensity of industrial noise, as well as the real-time requirements of industrial settings, indicate that none of the above algorithms are directly suitable for industrial human-computer interaction scenarios. Therefore, future research should focus on finding the different optimal weights to balance both high precision and lightweight for speech enhancement algorithms based on different types of industrial scenarios, and on constructing a hybrid enhancement system adapted to the noise characteristics of industrial scenarios

In conclusion, this review holds that the development of industrial speech interaction in the future should tend towards an integrated collaborative model, including perception, enhancement, recognition, and feedback layers. As shown in Figure 3, the "scene-noise type" label output by the perception layer triggers the adaptive switch of the enhancement layer algorithms. At the same time, the noise spectrum feature data is connected to the parameter adjustment interface of the enhancement module, which enables a dynamic mapping of the enhancement parameters if the noise characteristics are determined. Furthermore, the speech features output by the enhancement layer can be fused at the front end of the recognition layer (e.g., the ASR speech recognition model) with environmental features captured by the perception layer, such as noise types and device status, to improve the model's performance. In the feedback optimization stage, the recognition layer can respond hierarchically based on relevant thresholds (e.g., confidence scores). Simultaneously, the ternary data consisting of noise type, enhancement parameters, and recognition results from low-confidence scenarios are transmitted back to the enhancement layer to fine-tune the speech enhancement strategy, forming a closed loop of "collaboration - feedback - optimization."

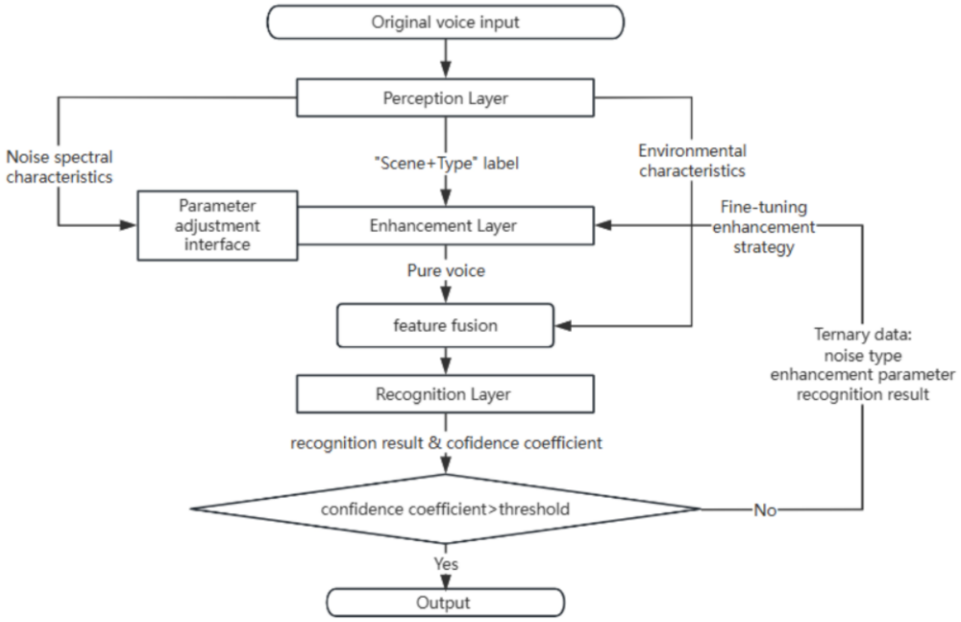

Fig. 3. Collaborative Integrated Industrial Speech Interaction Model. (Picture credit: Original)

5 Conclusion

This review aims to analyze the current developing states of industrial speech interaction. It analyzed the differences in multi-occupational environmental noise in various industrial scenarios and their impact on speech recognition models. It summarized the extant research, which focused on two core modules (speech denoising module and speech recognition module) in the industrial speech interaction process. At present, it has been proven that industrial multi-occupational noise reduces the accuracy of speech recognition models, and the impact on the performance of speech recognition models varies across different types of such noise. Deep learning-based speech enhancement algorithms (such as SEGAN and RNNoise) can efficiently learn the distinguishing characteristics between speech and noise, showing excellent adaptability and performance in noise suppression for multiple industrial occupations, and have become the mainstream research direction in this field. As the core subsequent stage in industrial speech interaction, the performance and noise adaptability of speech recognition algorithms directly determine the interaction effects. Different technical approaches, traditional methods versus deep learning methods, show distinct characteristics, and their performance can be greatly improved after optimization oriented by industrial adaptability. Equations and formulas should be punctuated in the same way as ordinary text but with a space before the punctuation mark.

In the future, multi-occupational noise types and industrial scenarios are expected to show a trend to be more complex. Constructing a collaborative integrated interaction system of perception, enhancement, recognition, and feedback layers may become a better research direction for industrial speech interaction. This review suggests that research on industrial speech interaction may develop along three pathways: noise-type and scenario mapping, personalized hybrid speech enhancement denoising models, and end-to-end integrated design for speech recognition.

References

[1] Wang, P.: Research and design of smart home speech recognition system based on deep learning. In 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL). pp. 218–221. IEEE, Chongqing, China (2020)

[2] Macháček, D., Dabre, R., Bojar, O.: Turning Whisper into real-time transcription system. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Bali, Indonesia (2023)

[3] Cognitive Market Research: Speech recognition software market report 2025 (Global edition). Cognitive Market Research (2025), https://www.cognitivemarketresearch.com/speech-recognition-software-market-report#request_sample

[4] Song, J., Chen, B., Jiang, K., Yang, M., Xiao, X.: The software system implementation of speech command recognizer under intensive background noise. IOP Conference Series: Materials Science and Engineering. Vol. 563, Art. no. 052090. IOP Publishing, Changsha, China (2019)

[5] Hamza, M., Khodadadi, T., Palaniappan, S.: A novel automatic voice recognition system based on text-independent in a noisy environment. International Journal of Electrical and Computer Engineering. 10(4), pp. 3643–3650 (2020)

[6] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning. Vol. 202. PMLR (2023)

[7] Rogowski, A.: Industrially oriented voice control system. Robotics and Computer-Integrated Manufacturing. 28(3), pp. 303–315 (2012)

[8] Li, S., Yerebakan, M. O., Luo, Y., Amaba, B., Swope, W., Hu, B.: The effect of different occupational background noises on voice recognition accuracy. Journal of Computing and Information Science in Engineering. 22(5), Art. no. 050905 (2022)

[9] National Institute for Occupational Safety and Health: Occupational hearing loss surveillance: Overall statistics – all U.S. industries. National Institute for Occupational Safety and Health (2025)

[10] Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: Dynamic noise aware training for speech enhancement based on deep neural networks. In Interspeech 2014. pp. 2670–2674. Singapore (2014)

[11] Pascual, S., Bonafonte, A., Serrà, J.: SEGAN: Speech enhancement generative adversarial network. In Proceedings of Interspeech 2017. pp. 3642–3646 (2017)

[12] Valin, J. M.: A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In IEEE Workshop on Multimedia Signal Processing. pp. 1–5. Vancouver, Canada (2018)

[13] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., et al.: Deep Speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the International Conference on Machine Learning. pp. 173–182. New York City, NY, USA (2016)

[14] Birch, B., Griffiths, C., Morgan, A.: Environmental effects on reliability and accuracy of MFCC based voice recognition for industrial human–robot interaction. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture. 235(12), pp. 1939–1948 (2021)
[15] Long, H., Huang, Z., Shao, Y., et al.: Research on language recognition algorithm based on improved CFCC feature extraction. Journal on Communications. 43(12), pp. 211–221 (2022)