Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **11**, Number **1**, February **2026**

Automatic Detection Method of a Multi-Language Recognition System Based on Artificial Intelligence Embedded Technology

Tingyu Luan[1], Xiaoguang Chen[1,*], Fengxia Zhang[1]

[1]College of Applied Technology, Dalian Ocean University,
Wafangdian, Dalian, 116300, Liaoning, China
15942420327@163.com, chenxiaoguang@dlou.edu.cn, 752791690@qq.com

Wei Liang[2]

[2]College of education, Valaya Alongkorn Rajabhat University, Bangkok 12120, Thailand
shixu8015@163.com

*Corresponding author: Xiaoguang Chen

Received May 08, 2025, revised August 20, 2025, accepted November 11, 2025.

Abstract. *With the rapid development of artificial intelligence technology, embedded systems are being increasingly used in the field of multilingual speech recognition, and research on their automatic error detection methods has become particularly important. This study proposes a multilingual speech signal denoising method based on wavelet transform and constructs an automatic detection mechanism to monitor the source and mode of recognition errors, improving the application stability of the system in complex environments. By comparing the application of intelligent and embedded technologies in multilingual speech recognition systems, their recognition accuracy, denoising performance and compatibility in different environments are analysed. The comparison information shows that the application of intelligent and embedded systems has increased by about 20%, and the recognition rate of embedded systems is 12% less than that of intelligent systems. Under the condition of 80 decibels, the discrepancy is reduced by 40%, which is a relatively large gap. Under the condition of 120 decibels, the gap between the two systems is about 50%. In terms of denoising, the gap between embedded and intelligent systems is about 10%. The compatibility of intelligent systems has increased by about 10%. Research results show that although embedded systems perform well in specific environments, intelligent speech recognition systems exhibit more advantages in high-noise environments. This study also discusses the calculation method of intelligent speech recognition systems and proposes a denoising method based on wavelet transform to improve the recognition accuracy of the system in complex environments. Finally, the establishment of a cross-compilation environment for embedded systems is discussed, providing technical support for the practical application of intelligent speech recognition systems.*
Keywords: artificial intelligence; embedded technology; multilingual speech recognition system; automatic error detection method; cross-compilation environment; wavelet transform.

1. **Introduction.** The growth of speech recognition technology is extremely convenient for the current lifestyle. Language is the easiest way to communicate. At present, in the era of "artificial intelligence," everything is interconnected, and everything is intelligent. Human–computer interaction is becoming increasingly frequent. Therefore, teaching a machine human skills that range from "hearing" to "understanding" is crucial [1, 2].

Speech recognition is instinctive for humans; for machines, however, it requires complex steps that involve input, computation, recognition, understanding and conversion into commands. Achieving human-level understanding remains a major challenge in artificial intelligence.

The first machine-based simultaneous interpretation occurred during the "21st Century Computing Conference" in Tianjin, China, wherein the potential of speech technology in tasks such as translation was highlighted. Speech recognition stands out amongst biometric technologies due to its ease of use, low cost and high acceptance, with less privacy concerns. However, it still faces limitations, including impersonation risks, remote misuse and susceptibility to factors, such as noise, emotions and signal distortion. Despite these challenges, intelligent systems can address many of the aforementioned issues, enhancing the accuracy and practicality of speech-based interactions. Artificial intelligence and speech recognition are illustrated in Figure 1.

Figure 1. Artificial intelligence and speech recognition

Although speech recognition systems have achieved significant progress in open language environments, existing research has mostly focused on recognition tasks under standard language resources, and the practical evaluation of embedded deployment in multilingual contexts is relatively insufficient. Simultaneously, most systems have not yet solved the misrecognition problem under high noise and strong interference conditions, and their automatic error detection mechanisms are also at a low level. The current study focuses on building a multilingual recognition system that runs on an embedded platform, combining wavelet transform for speech denoising and introducing an automatic recognition error detection strategy to improve the system's adaptability in low-resource and high-complexity scenarios.

2. **Related Work.** Multilingual speech recognition technology has developed rapidly in recent years. Researchers have generally focused on the construction of language models [3, 4], the improvement of acoustic modelling algorithms [5, 6] and the expansion of multilingual corpora to improve the adaptability of recognition systems in different language environments. With regard to end-to-end modelling, architectures based on attention mechanisms have been widely used in multilingual scenarios [7, 8]; these architectures have demonstrated good performance in reducing decoding complexity and improving speech recognition accuracy. However, most methods still rely on high-resource language data, exhibit insufficient generalisation capability under low-resource language conditions and are difficult to ensure deployment efficiency [9, 10].

In terms of the application of "artificial intelligence" algorithms, deep learning [11, 12] and machine learning [13, 14] are widely used for feature extraction and acoustic modelling, improving the robustness of the system to variant speech signals. In recent years, multitask learning [15, 16] and self-supervised learning [17] methods have gradually been introduced into speech recognition systems, improving the consistency of multilingual modelling. However, existing algorithms typically rely on high computing resources, and

thus, limitations still exist with regard to their adaptation to edge computing environments and embedded devices [18, 19].

The application of embedded platforms to speech recognition systems has expanded their low power consumption, light weight and mobile deployment capabilities, making them suitable for scenarios such as the Internet of Things and smart homes. At present, some studies have attempted to combine model compression and quantisation methods to reduce system complexity. However, most existing studies have focused on single-language recognition, whilst insufficient research has been conducted on system stability and recognition accuracy in multilingual environments. Problems, such as high misrecognition rate and frequent false triggering in embedded environments, have not been systematically solved.

In terms of error detection and recognition reliability, some scholars have proposed using confidence scoring mechanisms or error detection methods based on graph models to identify output anomalies [20, 21]. However, most of these methods are used as post-processing modules and lack coordinated optimisation with the speech signal modelling stage, resulting in system response delays and unstable detection accuracy [22].

In summary, the current multilingual speech recognition system on embedded platforms still exhibits evident deficiencies in real-time, adaptability and error control capabilities. Building an embedded recognition system for complex speech environments and combining it with an algorithm-level error detection mechanism provide a key technical path towards achieving high-performance multilingual recognition systems.

3. **Translation Method of the Embedded Intelligent Speech Recognition Input System.**

3.1. **Development Status of the Embedded Intelligent Speech Recognition System.**

From the perspective of the current development of intelligent voice, the present technology can be considered excellent. From the record in the 1921s, the sales volume of intelligent voice has reached nearly 30 billion yuan, which is highly evident from the data. The demand for intelligent voice is relatively strong and still in a state of continuous growth. The benefits of intelligent voice are considerable, and it is also a smart sense. The intelligent voice originated from the 1950s, but it has now entered the lives of common people. In the 1990s, computers were set up as a special research project every 2 years, i.e., basically the same as a foreign technology. Many well-known universities have conducted research on voice intelligence. At present, the accuracy of intelligent recognition can reach approximately 95%, with some cases of refusal to recognise. In English recognition, the recognition rate is nearly 95%. An embedded speech recognition system simply involves directly embedding the hardware into another hardware. Another possible meaning is to perform a superposition directly on the software. The two types of action objects are different: one is the system, and the other is the hardware. However, both actions can transform the intelligible into intelligent, which is their meaning. From the previous mechanical switches to voice control, light control, and finally, fingerprints, the development is real and can be observed, making people's lives increasingly simplified. This scenario demonstrates the appeal of technology. At present, voice has not yet fully entered people's lives. Elevators can be controlled by voice. Lights can be turned on by voice. Cooking can be done by voice. These directions represent subsequent development. This scenario is only for daily life, and many other fields are still available. In production, for example, voice control can simplify things by letting robots do dangerous things. The recognition algorithm for voice control has become numerous, and efficient recognition systems are

plenty. The speech recognition technology is a combination of many technologies, most of which are currently embedded. Its working principle is illustrated in Figure 2.

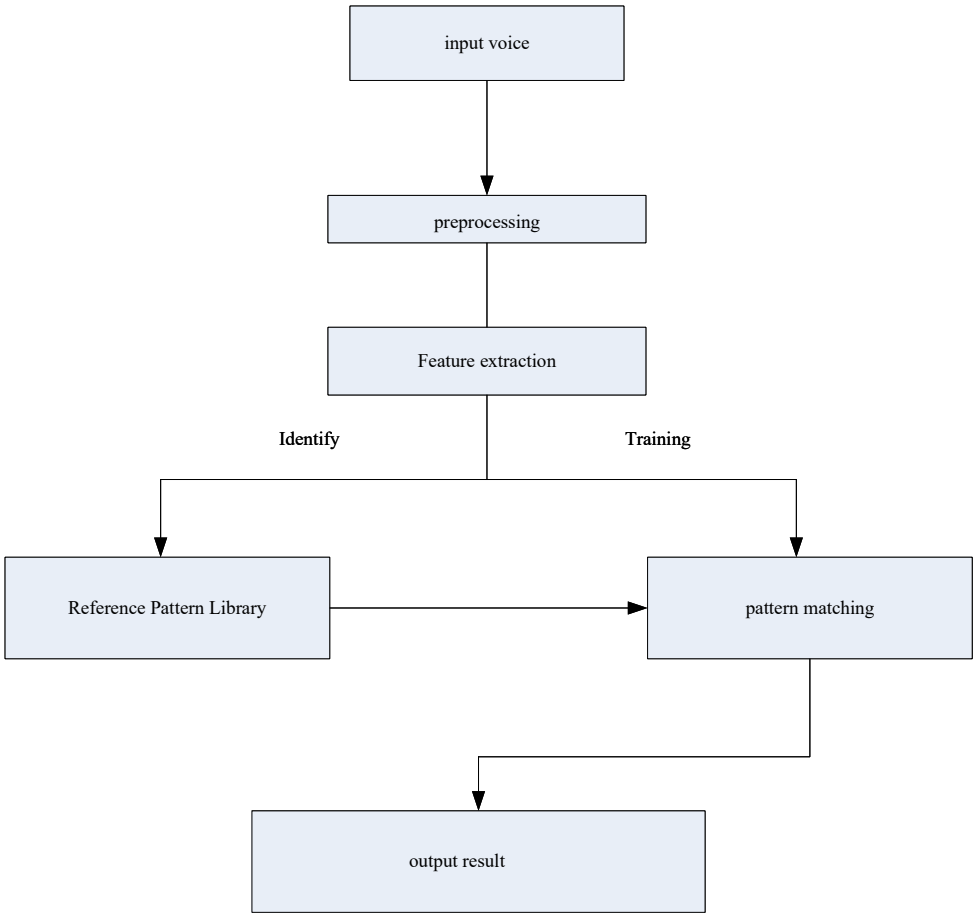

Figure 2. Schematic of speech recognition

At present, the recognition accuracy of instruments with speech recognition remains extremely high. With the embedding of smart chips, their corresponding technologies have been continuously improving. However, problems persist in some special cases. Compared with personal computers, embedded computers exhibit the advantages of small size, favourable price and portability [23, 24]. By the time the reaction and signal output end, embedded computers are still extremely fast and not slower than a personal computer. By contrast, they may even be faster, and their recognition accuracy is good. In addition, people's needs must be addressed in a timely manner, and embedded computers are highly advantageous in this aspect. With embedded computers, application identification can be performed anytime, anywhere, providing convenience to people's daily lives. The compatibility of embedded computers cannot be achieved by many other computer types. Embedded computers can be used with other functional systems, such as fingerprint unlocking and face payment. In terms of English word recognition, the recognition rate can reach 99%, whilst some specific dialects can also be recognised well. However, some unsolved problems persist. In terms of noise robustness, recognition accuracy is difficult in cases with considerable noise interference and challenging to achieve with the current technology. Meanwhile, a speech recognition system is primarily designed for single environments and single scenes due to its multi-category complexity, and performing general recognition in complex venues is difficult. In the MCB challenge, it wins the first place in the single item, whilst also achieving good results in several other items, such as

speech recognition, human segmentation and clustering, label alignment and asymptotic speech recognition. However, a problem exists in which not all types of languages in the world can be sorted into one voice system. This issue can finally be resolved in the era of artificial intelligence.

3.2. **Calculation Method of the Intelligent Speech Recognition System.** Firstly, the system recognises different sounds and tones. Then, identification is performed from the aspect of timbre, and intelligence achieves a higher degree of identification in this regard. An analysis of sound is conducted, and sound is converted into a signal. Then, corresponding to the pitch of the sound, it is converted into a signal with evident vocalisation, realising speech recognition. To clearly describe the specific implementation process of the model in an intelligent speech recognition system, the current study provides the algorithm logic in pseudo-code form:

Input:
x: Raw audio waveform

Output:
y_{pred}: Predicted language label

Procedure:
1: Apply pre-emphasis filter to x
2: Frame x into 25 ms windows with 10 ms step
3: **For each frame:**
  4: Compute Mel-frequency cepstral coefficients (MFCC) with 40 filter banks
5: Normalise all MFCC features across time
6: Feed normalised features into the neural network model
7: **Model architecture:**
  - Conv1D layer × 3
  - Bidirectional long short-term memory layer × 2
  - Fully connected classification layer
8: **For each frame:**
  9: Compute posterior probabilities over all language classes
10: Aggregate predictions over all frames via majority voting
11: Set $y_{pred} \leftarrow$ final predicted label

Return: y_{pred}

The mode diagram is shown in Figure 3.

Based on the principle of the figure, an analysis of the tone signal can be constructed. Assume that the sequence of the initial signal values is set to $x = [x(0) \dots x(N-1)]$. Amongst them, $x(n)$ is a finite number of speech analysis signals, and $n \in (0 \sim N-1)$. Then, x in the signal of the sound system is also called discrete wavelet transform after Fourier transformation. It can be expressed mathematically as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{N}nk\right) \tag{1}$$

Amongst them, $k \in (0 \sim N-1)$ represents the length of the signal. After signal $x(n)$ is changed, $X = DFT\{x\}$ can be used to represent the DFT of the finite sequence recognised by the speech system. Then,

$$X = [X(0) \sim X(N-1)] \tag{2}$$

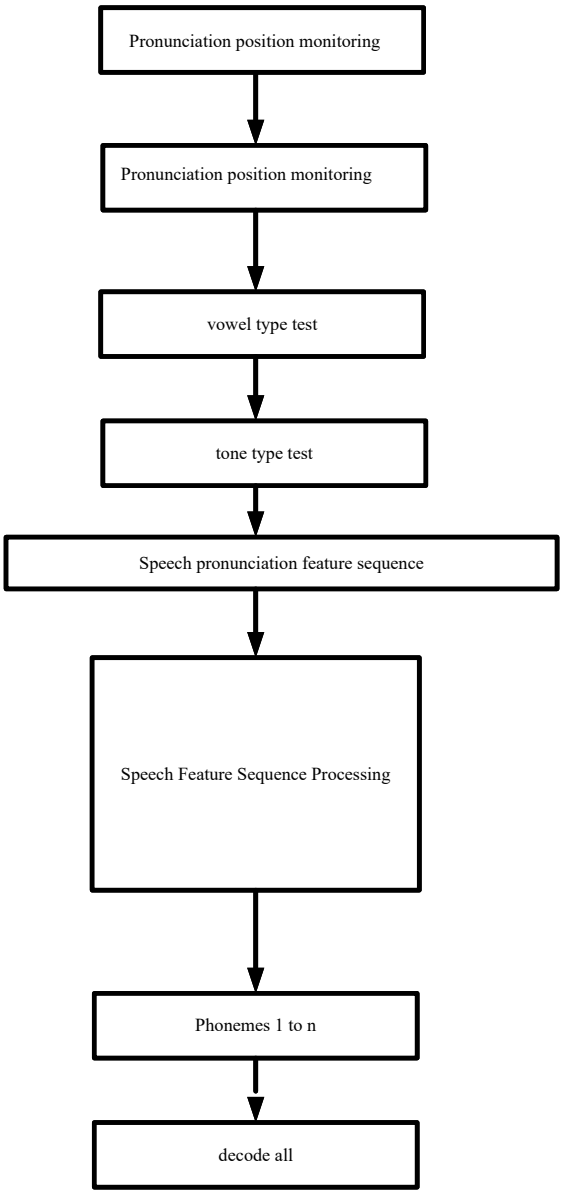

Figure 3. Schematic of the tone pattern

From the process illustrated in Figure 1, the wavelet signal inside is decomposed, and the voice obtained from it is adjusted to acquire a limited tone signal, which is rearranged and combined to form a voice signal with a different resolution of $j = 0, 1, 2 \ldots M$, i.e., $E_j = \sum_k |C_j(k)|^2$. The input signal for $N_0, 2N_1$ is $v_0(n), v_1(n)$, and its length is $C_j(k) = [x(t), \phi_j, k(t)]$. The two functions can be used to represent low-pass and high-pass wave transfers, $H_0(k), H_1(k)$, respectively. Then, the obtained signal output energy can be expressed as $E = \|x(t)\|^2 = \sum_j \sum_k |C_j(k)|^2 = \sum_j E_j$. The scale factor of the signal decomposition can be regarded as $N_0^j \approx \alpha^j N, N_1^j \approx \alpha^{j-1} \beta N$. Then,

$$N_0^j = 2round\left(\frac{\alpha^j}{2}N\right) \tag{3}$$

$$N_1^j = 2round\left(\frac{\alpha^{j-1}}{2}\beta N\right) \tag{4}$$

Amongst them, $j \in (1 \sim J)$ represents the amplitude coefficient of wavelet transform. In accordance with the equation, some desired waves are automatically selected via filtering, and some unwanted clutters are discarded.

This part is about the filtering process and the process for the information contained in it. After selecting the layers of the input wave information, the intelligent voice system is entered. The signal parameter in the beginning is assumed as $N^{(j)}$, which is the length of a signal on the j-layer wave filter. N_0^j, N_1^j represent the length the pronunciation in the input system. P_j represents an energy set in the signal region, and it can be the distribution coefficient of $\{P_1, P_2 \dots P_j\}$ in the wavelet. In the wavelet band $\omega^{(j)}$, the length is $N_1^{(j)}$, and the same frequency band between the signals is $C^{(j)}$. Therefore, some of its filter coefficients can be obtained, which are 4, $N^1 = N, N^j = N_0^{j-1}$, and $j \in (2 \sim J)$. Each of its signals is represented by P_j, because it can be deduced to $\sum_j P_j = 1$. Then, the obtained signal is subjected to a noise reduction process in accordance with the wavelet band decomposition method. Noise reduction processing is performed on an input signal value through function decomposition, and certain clear audio information can be obtained.

After using the wavelet function decomposition method, the intelligent voice is inputted into the system, and different voice and tone signals are calculated using an automatic algorithm under a decluttering process to obtain a piece of information. After decomposing the wavelet and picking up a clearer intelligent speech based on its characteristics, the unstable fluctuations in it are changed and then automatically recognised, and the speech of each frame is analysed. Assume that the number of frames is Z_n. Then, the distribution of the number of frames on the audio is:

$$Z_n = \sum_{-\infty}^{\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|\omega(n-m) \tag{5}$$

In the equation,

$$sgn[x] = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{6}$$

$$\omega(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & \text{else} \end{cases} \tag{7}$$

The wave signal energy obtained by each layer is rearranged and combined to obtain a new signal component, which can be expressed as E_n.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\omega(n-m)]^2 \tag{8}$$

The coefficients of the wavelet signal of each segment are:

$$T_j = \begin{cases} \sigma\sqrt{2\ln(N)}\left(1 - \frac{J}{2} \times \frac{E_j}{\sum_{j=1}^{J+1} E_j}\right) & j = 1, 2, 3, \dots J \\ \sigma\sqrt{2\ln(N)} \times \frac{E_j}{\sum_{j=1}^{J+1} E_j} & j = J+1 \end{cases} \tag{9}$$

where N represents the length of a signal; and J represents the frequency of a sound wave, which can be determined by a new threshold as $\omega_j^k(X)$.

$$\omega_j^k(X) = \begin{cases} sign\,\omega_j^k(|\omega_j^k| - \beta \cdot T_j) & \text{if } |\omega_j^k| \geq T_j \\ 0 & \text{else} \end{cases} \tag{10}$$

where β is the adjustment meson of the sound, and its value is between 0 and 1. The wavelet entropy value after extracting the speech is:

$$T_j = \sigma\sqrt{2\ln(N)}, \quad j \in (1 \sim J+1) \tag{11}$$

$$\sigma_j = \frac{median(d_j(k))}{0.6745} \tag{12}$$

where $d_j(k)$ represents the Euclidean distance, which is the straight-line distance between two points in 2D or 3D space. The hard interval function for adjusting the pitch of the input system can be obtained:

$$\omega_j^k(X) = \begin{cases} \omega_j^k & \text{if } |\omega_j^k| \geq T_j, j \in (1 \sim J+1) \\ 0 & \text{else} \end{cases} \tag{13}$$

Then, the obtained soft interval function is:

$$\omega_j^k(X) = \begin{cases} sign(\omega_j^k)(|\omega_j^k| - T_j) & \text{if } |\omega_j^k| \geq T_j \\ 0 & \text{else} \end{cases} \tag{14}$$

The following part describes how the instrument learns. Assume that the training samples of the machine are X and Y. The value is predicted as accurately as possible, as shown in Figure 4.

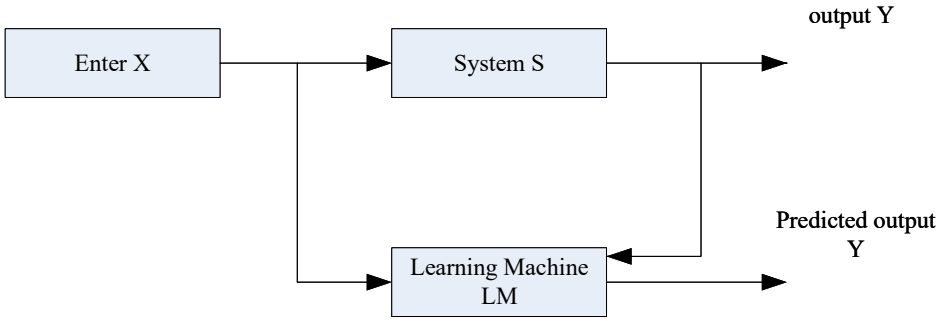

Figure 4. Relational block diagram for machine learning

The input X is an arbitrary variable that affects the value of $F(x)$ of the distribution function. S is a trainer of the system. A randomly occurring Y is obtained in accordance with different random variables. In accordance with the distribution function $F(Y/X)$, a conclusion can be drawn that random values are extracted from the function to obtain a large separate training set, such that the new function can be closer to the output function. This set of functions is also called the prediction function. In practical applications, the training sample X can be represented as a set of feature parameters of speech segments in different languages, and Y is the language label category outputted by the system. In multi-language recognition tasks, the system must determine language affiliation based on the acoustic features of the input to match the corresponding recognition model. For example, the same ambiguous speech sample is inputted into the prediction function set obtained by training multiple language models, and finally, the optimal recognition path is selected based on the principle of minimum error.

Compared with the distribution function, this prediction function is more accurate. The obtained set of functions can be $f(x, a)(a \in A, A$ is a set of parameters), and the

best set of functions $f(x, a^*)$ can be selected to minimise the expected risk, which can be presented as follows:

$$R(a) = \int L(y, f(x, a))dF(x, y) \tag{15}$$

Amongst them, $f(x, a)$ can be called a function set of prediction. $L(y, f(x, a))$ is the loss expectation given by the system output and the learning machine. Different function types can be used to form different term machines. The definition provided in Figure 3 is relatively simple, and many meanings can be represented by mathematical expressions. In general, the major problems of machine learning are generally to classify the types of recognition and the infinitely close linear regression problem in the function risk [25]. The issue of minimising the estimated risk also exists. The following part provides an overview of speech recognition, whilst the two other ? are not explained. In the problem of classifying the output function, the output y can be defined as 0 or 1, and the prediction function set is called the indicator function set in this paper. Then, the instantaneous function set can be:

$$L(y, f(x, a)) = \begin{cases} 0 & y = f(x, a) \\ 1 & y \neq f(x, a) \end{cases} \tag{16}$$

When the value is 1, the obtained function is classified as a dislocation set, and a risk functional can also be derived for this data set to determine the accuracy of the prediction function and the probability obtained by the trainer. By comparing multiple sets of data, it is also called the mean error rate [26]. The primary purpose of machine learning is to obtain the accuracy of (23), which can only be realised by comparing multiple sets of data. Certain empirical risks also exist. The following mathematical expressions can be used:

$$R_{emp}(a) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(x_i, a)) \tag{17}$$

The expected risk is small, but it does not mean that the empirical risk is small. The value of the sample can be infinitely close to infinity, but it cannot be infinite in the collection of speech classes, and a new method is required to achieve it.

3.3. **Establishment of the Embedded System's Cross-Compilation Environment.** The embedded system is a closed environment that is similar to a ring, and it is relatively difficult to test directly. The programme is also more complicated and typically not feasible [27, 28, 29]. Most of the methods that have been used at present are for writing the data programme on the host side firstly, running the binary directly on the hardware through cross-compilation and finally downloading it directly to the development board by using some methods. Cross-compilation means that the corresponding programming can be generated, read and copied on another platform. It also has two meanings [30]. That is, the same system can run different systems, and vice versa, and systems can run back and forth with one another. The cross-compilation process is illustrated in Figure 5.

4. Comparative Analysis of Intelligent and Embedded Speech Recognition.

4.1. **Comparative Analysis of Users.** As people's lives have been considerably improved, an increasing number of people have begun to understand intelligent products. From voice recognition in a car system to "Xiao Ai" on a mobile phone, to a mobile phone's intelligent navigation system, all of which have considerably facilitated people's lives. In terms of speech, however, what is the difference between embedded and intelligent systems? In the past 10 years, this technology has been developing steadily. One

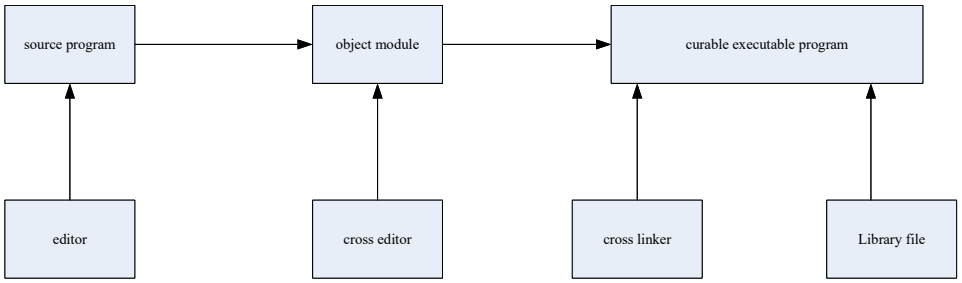

Figure 5. Schematic of cross-process compilation

experiment conducted a comparative analysis of different groups and types. Amongst the 400 users who participated in the test, 8% were under 18 years old, 35% were between 18 years and 30 years, 38% were between 31 years and 50 years and 19% were over 51 years old. In terms of occupation, students accounted for 32%, teachers accounted for 17%, corporate employees accounted for 22% and freelancers and retirees accounted for 29%. These data are provided in Tables 1 and 2.

Table 1. Number of users of embedded identification system

	Students	Teachers	Workers	Elders
Number of users	53	48	60	9
Number of users	56	45	61	10
Number of users	58	43	56	12
Number of users	60	42	58	13

Table 2. Number of users of intelligent speech recognition system

	Students	Teachers	Workers	Elders
Number of users	86	65	75	38
Number of users	84	64	74	36
Number of users	85	67	74	34
Number of users	81	69	72	31

By comparison, more people use an intelligent voice system, and it is more common today. At the student level, the data used have increased by nearly 50%. The number of teachers is similar. The number of ordinary workers has also increased by about 20%. Improvement is the largest at the elderly level, because the elderly cannot use complicated items, and an intelligent voice system is more convenient. Use amongst the elderly has increased by nearly four times. These findings can also explain the convenient advantages of using intelligent speech recognition.

4.2. **Comparative Analysis of Recognition Accuracy.** In a certain degree of comparison, ensuring that other influences are consistent is necessary. Multiple tests in the same environment are selected to determine the accuracy of the two systems and how the two systems compare in different environments, as depicted in Figure 6.

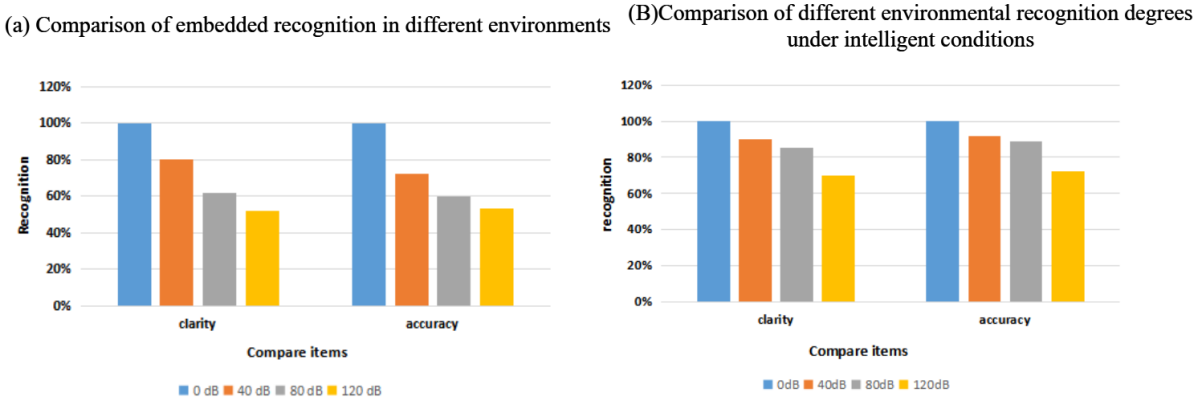

Figure 6. Comparison between embedded and intelligent voice recognition

Through a certain comparison, in the case of a quiet environment, not much difference is found between the two systems, but a gap exists in the follow-up. In the same environment of 40 decibels, the embedded recognition rate is 12% lower than that of the intelligent one. In the case of 80 decibels, it is relatively reduced by 40%. This difference is huge. In the case of 120 decibels, the difference between the two is about 50%. In the case of noisy sound, the gap between embedded and intelligent speech recognition is increasing. These findings shows that the intelligent type exhibits more advantages in the current environment. It can also recognised better in a noisy environment, and it is more developmental in the current times.

4.3. **Comparison and Analysis on Noise Removal.** At present, people encounter noise as they live and work. Noise can only be avoided as much as possible, but no method is available to reduce it. Everyone basically has things that must be sorted out in time, and thus, a relatively quiet environment is necessary and more conducive to adapting to a new environment. A judgment is made on the comparison in noise removal performance between the two systems at present, as illustrated in Figure 7.

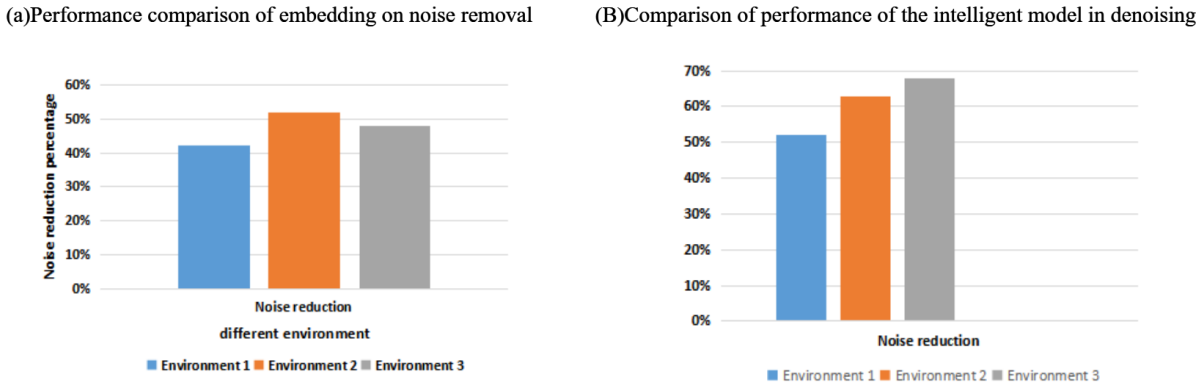

Figure 7. Comparison of noise removal between the embedded and intelligent speech systems

Through a comparison of the three sets of experiments in embedded noise removal, the difference between the embedded and intelligent models is about 10%, demonstrating evident contrast in follow-up life. The current era is also moving forward gradually; hence, people also choose a better living environment. The operation is simpler for intelligent individuals, and it is more suitable for people of different ages.

4.4. **Compatibility Comparison.** With the update in modern hardware and software, people's choices are more favourable. Many incompatible software and hardware are gradually eliminated after they cannot be used in the follow-up, and thus, the current study conducts experiments and comparisons for compatibility selection to determine the practicability and compatibility of hardware and software in various industries. The embedded platform used in this test is Raspberry Pi 4B based on the ARM Cortex-A53 architecture, equipped with a 64-bit Debian system and OpenCV 4.5.1 and TensorFlow Lite 2.8 versions. The intelligent speech recognition test environment is a laptop with an Intel i5-1135G7 processor and Windows 10 operating system, with standard TensorFlow 2.8 and PyTorch 1.10 platforms. The same speech data set is used for model deployment and recognition testing in the software environment to ensure consistent test conditions and improve the comparability of compatibility evaluation, as depicted in Figure 8.

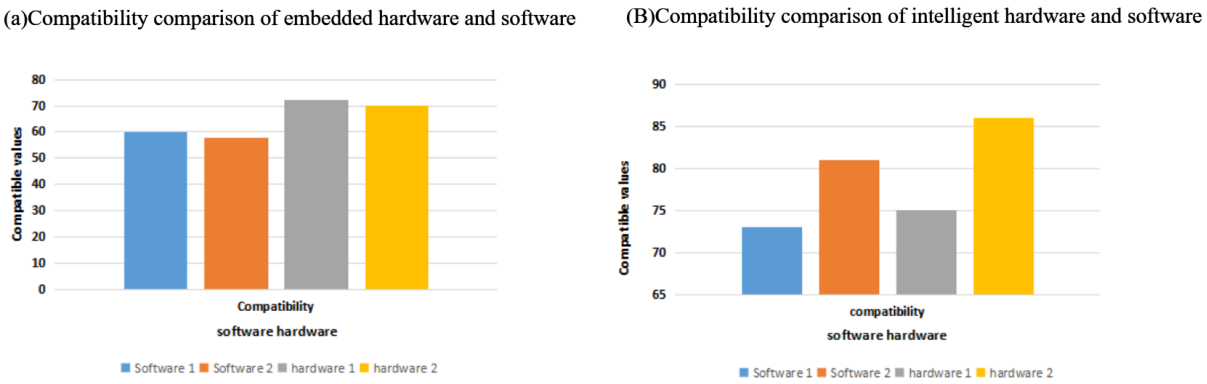

Figure 8. Compatibility comparison between embedded and intelligent voice recognition

Through a comparison of compatibility, compatibility in software is improved by about 15% compared with the embedded model, whilst compatibility in hardware is improved by about 10%. The same result can still be noted in the comparison of the same type of hardware and software. The practical application of the two platforms is also gradually increasing in intelligence, demonstrating an upward trend, and follow-up may be eliminated one after another. At present, many fields are using intelligent technology.

5. Discussion.

5.1. **Convenience and Application Brought by Intelligent Speech Recognition in the Age of Artificial Intelligence.** With the rapid development of science and technology, people are increasingly using technology quickly and conveniently, gradually simplifying people's lives. In terms of communication, information can be broadcasted directly through TV stations or written and published on newspapers, which can be read directly. After a conversion, this information becomes needed by the people. With the help of "artificial intelligence," many apps, such as novels, listening books and news, can be simplified. In this case, people can speak freely and express their own ideas and opinions. Live broadcasts are also available, which people can send as they want, enhancing communication amongst individuals. For example, phone calls reduce the cost of communication and make it more convenient. Many people also directly communicate through WeChat voice. After all, voice is the easiest way to communicate, which has been the situation for thousands of years.

5.2. **Possible Disadvantages of Intelligent Speech Recognition System.** Firstly, people may become increasingly dependent on "artificial intelligence" voice systems, become lazy and refuse to think on their own. Speech recognition should be within a certain distance range. The current speech recognition system faces interference from many external factors, such as echo, reverberation, human voice, geographical and other factors. In addition, many types of languages, such as dialects, exist, leading to the low accuracy of speech recognition and weakening contextual connection, making learning and building models difficult. Moreover, this situation can easily lead to an increase in the unemployment rate of subsequent employees. Robots do not get tired and do not require wages, exerting pressure on some employees in the assembly line with corresponding tasks. From a macro perspective, people may lose emotions in the follow-up. Supposing that for the rest of their life, people's friends are emotionless robots, leading to the deterioration of life skills. In terms of danger, if the wrong people use this technology, then the danger posed to people is considerable.

In a high-noise environment, the signal front end of the intelligent speech recognition system is easily disturbed by unstructured background sounds, and the signal-to-noise ratio of the input speech signal drops sharply, resulting in distortion in the feature extraction stage, which, in turn, affects the stability of model prediction. The current mainstream speech recognition model relies on specific speech corpus during the training process and lacks generalisation capability for unknown noise types. System recognition output is prone to producing wrong labels. Such problems are more serious in complex background environments, such as factories, transportation hubs and public venues. Although noise reduction algorithms and speech enhancement technologies have improved the robustness of the system to a certain extent, the system still experiences problems, such as recognition interruption, command mismatch and response delay, when multiple sound sources exist simultaneously or noise power is greater than speech power, seriously affecting the interactive experience. Therefore, the semantic preservation mechanism and end-to-end noise modelling structure in high-noise scenarios still require further exploration.

6. **Conclusions.** This study compares between embedded and intelligent speech recognition and translation systems, and introduces the initial development of speech recognition systems to the present development. Some social problems may result from the use of speech recognition in contemporary times for subsequent convenience. This study also explores the algorithm for this technology, and to a certain extent, considers some possible negative aspects. A brief statement of the identification process of the technology is presented. A wavelet signal input into the speech system is decomposed, rearranged and combined into a new combination. In a large set, it is divided into two extreme waves, low-pass and high-pass transfer waves. After wavelet transformation, several required waves are filtered. After filtering, a noise reduction process is also performed on input information. Through function decomposition, an adaptive wave frequency is obtained. The original signal is debugged through a filter to obtain the result value. The information of each segment is decomposed on the wavelet. After filtering, a relatively clear speech is obtained. Some unstable fluctuations occur. Here, a number of frames is required, and a clear sound can be obtained by judging the threshold of the number of frames in the audio frequency. This study also introduces how instruments learn from one another mostly through repeated tests to obtain a correct set of functions to judge the accuracy and probability of the learning ability predicted by the system. It also introduces the environment in embedded cross-compilation, and compared the intelligent and embedded platforms in various aspects to reduce identification error. In terms of artificial intelligence, however, follow-up development is definitely huge and has applications in many

aspects. This situation is also a double-edged sword, depending on how these technologies are used. Speech recognition has greatly facilitated our lives, and future life is expected to become more convenient in terms of technology. Future research can focus on optimising the real-time and energy-saving performance of multilingual speech recognition systems in edge computing environments and on improving the system's dynamic adaptability in multitask and multi-language contexts. Simultaneously, in terms of automatic error detection, a self-supervised learning mechanism can be introduced to improve the model's recognition accuracy for low-resource languages. In view of differences in speech features amongst various languages, building a unified multilingual speech embedding representation system is also a key technical direction towards improving the versatility of the system.

REFERENCES

[1] H. Fan, "Research on innovation and application of 5G using artificial intelligence-based image and speech recognition technologies," *Journal of King Saud University-Science*, vol. 35, no. 4, pp. 102626–102634, 2023.

[2] J. L. K. E. Fendji, D. C. M. Tala, and B. O. Yenke, "Automatic speech recognition using limited vocabulary: A survey," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 2095039–2095075, 2022.

[3] Y. Zhao, W. Zhang, and G. Chen, "How do large language models handle multilingualism?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 15296–15319, 2024.

[4] K. Shanmugavadivel, V. E. Sathishkumar, and S. Raja, "Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data," *Scientific Reports*, vol. 12, no. 1, p. 21557, 2022.

[5] A. Singh, N. Kaur, and V. Kukreja, "Computational intelligence in processing of speech acoustics: a survey," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2623–2661, 2022.

[6] A. Sruthi, A. K. Kumar, and K. Dasari, "Multi-language: ensemble learning-based speech emotion recognition," *International Journal of Data Science and Analytics*, vol. 19, no. 3, pp. 453–467, 2025.

[7] A. Fateh, R. T. Birgani, and M. Fateh, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," *IEEE Access*, vol. 12, pp. 41381–41395, 2024.

[8] M. K. Majhi and S. K. Saha, "An automatic speech recognition system in Odia language using attention mechanism and data augmentation," *International Journal of Speech Technology*, vol. 27, no. 3, pp. 717–728, 2024.

[9] J. Bai, "Design of the artificial intelligence vocal system for music education by using speech recognition simulation," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 5066004–5066013, 2022.

[10] J. Liu, K. Li, and A. Zhu, "Application of deep learning-based natural language processing in multilingual sentiment analysis," *Mediterranean Journal of Basic and Applied Sciences (MJBAS)*, vol. 8, no. 2, pp. 243–260, 2024.

[11] R. Geethanjali and A. Valarmathi, "A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media," *Scientific Reports*, vol. 14, no. 1, pp. 22270–22297, 2024.

[12] T. Harinadh, M. A. Valli, and A. V. Chowdary, "Enhancing cross-language understanding: A machine learning-based approach to multilingual identification," *International Journal for Modern Trends in Science and Technology*, vol. 11, no. 03, pp. 418–428, 2025.

[13] A. Fateh, R. T. Birgani, and M. Fateh, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," *IEEE Access*, vol. 12, pp. 41381–41395, 2024.

[14] R. Imaizumi, R. Masumura, and S. Shiota, "End-to-end Japanese multi-dialect speech recognition and dialect identification with multi-task learning," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, pp. e4–e27, 2022.

[15] L. Yunxiang and Z. Kexin, "Design of efficient speech emotion recognition based on multi-task learning," *IEEE Access*, vol. 11, pp. 5528–5537, 2023.

[16] R. Jain, A. Barcovschi, and M. Y. Yiwere, "A Wav2Vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46938–46948, 2023.

[17] J. Dong, "Natural language processing pretraining language model for computer intelligent recognition technology," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 8, pp. 1–12, 2024.

[18] X. Huang, D. Zou, and G. Cheng, "Trends, research issues and applications of artificial intelligence in language education," *Educational Technology & Society*, vol. 26, no. 1, pp. 112–131, 2023.

[19] X. Zhu, J. Li, and Y. Liu, "A survey on model compression for large language models," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1556–1577, 2024.

[20] A. S. Dhanjal and W. Singh, "A comprehensive survey on automatic speech recognition using neural networks," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23367–23412, 2024.

[21] J. Tobin, P. Nelson, and B. MacDonald, "Automatic speech recognition of conversational speech in individuals with disordered speech," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4176–4185, 2024.

[22] Q. B. Diep, H. Y. Phan, and T. C. Truong, "Crossmixed convolutional neural network for digital speech recognition," *PLOS ONE*, vol. 19, no. 4, pp. e0302394–e0302416, 2024.

[23] J. Chen, T. H. Teo, and C. L. Kok, "A novel single-word speech recognition on embedded systems using a convolution neuron network with improved out-of-distribution detection," *Electronics*, vol. 13, no. 3, pp. 530–546, 2024.

[24] W. He, "Automatic error detection method of embedded English speech teaching recognition system under the background of artificial intelligence," *Mobile Information Systems*, vol. 2022, no. 1, pp. 7340051–7340062, 2022.

[25] E. Liu, K. Chen, and Z. Xiang, "Conductive particle detection via deep learning for ACF bonding in TFT-LCD manufacturing," *Journal of Intelligent Manufacturing*, vol. 31, no. 4, pp. 1037–1049, 2020.

[26] U. Zabit, K. Shaheen, and M. Naveed, "Automatic detection of multi-modality in self-mixing interferometer," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9195–9202, 2018.

[27] M. Lv, "Agricultural climate change and multilingual GIS database translation system based on embedded database and artificial intelligence," *Arabian Journal of Geosciences*, vol. 14, no. 11, pp. 1–20, 2021.

[28] R. Kanniga Devi, M. Gurusamy, and P. Vijayakumar, "An efficient cloud data center allocation to the source of requests," *Journal of Organizational and End User Computing (JOEUC)*, vol. 32, no. 3, pp. 23–36, 2020.

[29] D. N. Kanellopoulos, "Congestion control for NDN-based MANETs: Recent advances, enabling technologies, and open challenges," *Journal of Organizational and End User Computing (JOEUC)*, vol. 33, no. 5, pp. 111–134, 2021.

[30] J. Vavrek, P. Viszlay, and M. Lojka, "Weighted fast sequential DTW for multilingual audio query-by-example retrieval," *Journal of Intelligent Information Systems*, vol. 51, no. 2, pp. 439–455, 2018.