Key Information Recognition in Speech Using Spectral Subtraction and Wavelet Thresholding Methods: An Optimized Approach with the S-PaddleSpeech Model

Yang Yang
Department of Information Management, Henan Vocational College of Light Industry, Zhengzhou 450008, China
E-mail: yang_yangyy@outlook.com

Keywords: speech recognition, paddlespeech model, keyword detection, spectral subtraction, wavelet threshold denoising method

Received: November 14, 2024

A PaddleSpeech model was developed to address the issue of low accuracy in speech key information detection technology. Spectral subtraction is used in the process to enhance the quality of speech signals and improve signal-to-noise ratio by reducing noise interference. The method of signal processing through wavelet transforms balances the selection of appropriate denoising methods in both time and frequency domains. Frame segmentation, windowing, and Fourier transform techniques were used in the data processing stage. The experiment outcomes show that for specific and non-specific speech, the CNN detection algorithm achieves a keyword recognition accuracy of 0.9 when the sample size is less than 20, while the FNN algorithm achieves an accuracy of 0.8 when the sample size reaches 60. Both in terms of sample size requirements and keyword recognition accuracy, CNN outperforms FNN. In addition, in the application testing of the model, the improved PaddleSpeech model shows significantly better recognition performance for 20 keywords in audio than the original PaddleSpeech model, with a recognition accuracy of up to 90% (P<0.05). In the audio character recognition verification of the improved PaddleSpeech model and SpeechRecognition model, the former correctly recognizes 15019 characters with an accuracy of 98.9580%, while the latter correctly recognizes 14593 characters with an accuracy of 96.1520. The former has an accuracy 2.806% higher than the latter (P<0.05). Therefore, the improved PaddleSpeech model proposed by the research has good speech keyword recognition ability and effectively improves recognition accuracy.

Povzetek: Odprtokodni orodni komplet za obdelavo govora in zvoka PaddleSpeecch je izboljšan z metodo spektralne subtrakcije in CNN algoritmom, kar doseže višjo robustnost v hrupu ter kvaliteti prepoznavanja.

1 Introduction

Due to the swift advancement of mobile Internet, computer technology has found extensive application across numerous societal sectors, with which a large amount of data of various types is produced rapidly [1]. Among them, due to the fact that voice is the most direct and effective method for people to engage in social activities, modern intelligent devices are constantly increasing their voice interaction functions, so voice data has received more and more attention from people [2]. To process various speech data, various speech recognition technologies are constantly being updated and upgraded. Among them, in the process of voice interaction between smart devices and humans, the key information recognition technology of speech performs well in various speech recognition scenarios due to its high computing power and fast recognition speed [3, 4]. However, there is still an issue of inadequate identification precision in current speech recognition systems when autonomously recognizing speech content [5]. On the one hand, when recognizing a large number of words simultaneously, speech recognition systems may experience spelling errors and omissions, thereby reducing accuracy. Studies indicate that, in environments where over 150 words are processed at once, error rates may spike by up to 35%, leading to a drop in recognition accuracy to below 65%. This is particularly problematic in real-time applications where precision is crucial [6]. On the other hand, the system's performance can suffer from significant keyword omission due to the extensive vocabulary and prolonged speech recognition time. Research has shown that in scenarios involving lengthy dialogues — typically exceeding 300 words — the occurrence of keyword omissions can reach as high as 25%. This not only impacts the system’s ability to capture essential information but also reduces the effectiveness of extracting key information in critical applications, such as automated transcription or real-time translation [7, 8]. Therefore, the PaddleSpeech model is constructed and improved using speech enhancement algorithms and keyword detection algorithms to raise the detection capability of speech and the precision of key content recognition.

The study is organized into four parts. The first part presents an overview and analysis of the domestic and international research on speech key information recognition technology. The second part is to construct the PaddleSpeech model and optimize it. The third part is to testify the general capability of the algorithm and the improved PaddleSpeech model. The last part is an overview of the whole piece of writing.

2 Related work

With the progress of computers and networks, an increasing number of artificial intelligence (AI) techniques are being utilized to various aspects of life. Voice interaction system is one of them. With the support of speech recognition technology, people can communicate with each other and machines through smart devices. Faced with the increasing demand for voice interaction, how to improve the accuracy of voice detection has become a growing concern for researchers. Yuan Q et al. explored English speech translation recognition technology based on long short-term memory network algorithm for speech information recognition problem. The study analyzed the shortcomings of traditional template matching and statistical pattern recognition, and proposed the application of long short-term memory networks to achieve intelligent information processing by simulating the information processing of the human brain. The results showed that the speech recognition accuracy of the long short-term memory network was as high as 94%, significantly improving information storage efficiency and increasing the processing speed of speech data to a maximum of 4.5 seconds, effectively meeting the society's demand for intelligence [9]. Xu et al. proposed a novel bimodal emotion recognition algorithm for the problem of emotion recognition in speech information recognition. This algorithm combines parallel convolution modules and attention based bidirectional long short-term memory modules to achieve feature abstraction and fusion. The experimental results show that the recognition accuracy of audio data reaches 74.70%, text is 77.13%, and the accuracy of the bimodal fusion model is as high as 90.02%. This study demonstrates the feasibility of processing heterogeneous information within homogeneous network components, providing flexibility for modal extension and architecture design [10]. For the problem of automatic speech recognition, Zhang et al. proposed a method called BigSSL, which used a giant Conformer model pre-trained with wav2vec 2.0 and combined SpecAugment technology for noisy student training. This method reached a word error rate of 1.4%/2.6% on the LibriSpeech test set and other test sets, which was a breakthrough in comparison with the cutting-edge WER of 1.7%/3.3% at that time [11]. To handle the matter of low accuracy in speech emotion detection, Zhu et al. first proposed two baseline methods, namely Vanilla Fine-tuning (V-FT) and Task Adaptive Pre-training (TAPT), to evaluate the capability of Wav2Vec 2.0 in SER. The experiment outcomes showed that on the IEMOCAP dataset, V-FT could surpass existing cutting-edge models, demonstrating the potential of Wav2Vec 2.0 in SER tasks [12]. Ambrogio et al. raised a simulated AI chip to handle the matter of low energy efficiency of AI models in high-precision tasks. Researchers have introduced a new type of simulated AI chip that integrates 35 million phase change memory devices on 34 tiles, large-scale parallel inter tile communication, and simulated low-power peripheral circuits, achieving chip sustained performance of 12.4 megaoperations per watt per second [13].

In addition, Burchi M and Timofte R proposed an efficient Conformer model for audio visual fusion to address the performance degradation of automatic voice recognition technologies in loud settings. Recently, although end-to-end ASR systems grounded on neural networks have shown excellent performance on clean audio samples, their performance often deteriorates under noisy conditions. To resolve this problem, scholars developed a method that combined audio and visual modalities to enhance the noise robustness of an efficient conformal connected temporal classification architecture. They relaxed the conditional independence assumption based on the CTC model by introducing residual modules between CTCs, and replaced efficient consistency grouping attention with a more efficient and simpler patch attention mechanism [14]. Deng et al. proposed a position guidance process monitoring and vibration detection method based on smartphones using PaddleSpeech for speech information recognition problems. In the study, smartphones were used to record the human-machine interface and tool movements of machine tools, and the correlation between tool cutting positions and sound signals was obtained through optical character recognition technology. Using PaddleSpeech's open-source model for speech recognition and voice separation, removing periodic components, and using the ratio of residual signal energy to total signal energy for vibration detection. Finally, this method was validated through robot milling and deep hole drilling experiments [15]. In response to the issue of long-term companionship between humans and pets, Chen Jiancheng et al. used Autoencoder and PaddleSpeech models, combined with VR systems and deepfake technology, to restore the appearance of family members and alleviate users' longing for their loved ones [16]. Wu et al. used PaddleSpeech to construct a corpus for speech information recognition, which is the first natural audio-visual multimodal database for Chinese social interaction agents. It includes 48 hours of videos and annotations, covering eight modalities. The study analyzed the characteristics of voice, language, behavior, and multimodal combinations during the questioning process, and tested the performance of six baseline models on three tasks. The results of this project provide reference for designing social interaction agents that are more in line with Chinese culture and user needs, and promote the improvement of Chinese daily data processing [17]. As a reaction to the problem that current depression detection models were difficult to represent small changes in depression, Zhou et al. raised a multi-granularity fusion network, which integrated different feature information to promote multi-level information and multi-resolution interaction, thereby optimizing the effectiveness of depression detection [18]. The summary and analysis of related work are shown in Table 1.

Table 1: Summary and analysis of related work

Reference	Advantages	Disadvantages	Performance Metrics	Advantages of This Study's Method
[7] Yuan et al.	High accuracy in English speech recognition	May not be suitable for non-English speech	Recognition accuracy: 94%	Better for Chinese speech recognition
[8] Xu et al.	Accurate in dual-modal emotion recognition	Requires textual information	Audio recognition accuracy: 74.70%	More suitable for pure speech recognition
[9] Zhang et al.	Good performance in large model pre-training	High computational resource requirements	Word error rate: 1.4%/2.6%	High accuracy with small samples
[10] Zhu et al.	High potential with Wav2Vec 2.0 and large data	Requires domain-specific data	Error rate: Lower than existing models	Good performance across multiple fields
[11] Ambrogio et al.	High efficiency of AI chip	Technical challenges in actual deployment	Energy efficiency: 12.4 megaoperations per watt per second	Easier to promote with software optimization
[12] Burchi and Timofte R	Improved robustness in noisy environments	Requires visual information	Robustness improvement: Significant	Applicable without visual information
[13] Deng et al.	Smartphone monitoring and vibration detection	Dependent on specific hardware	Recognition accuracy: High	Not dependent on specific hardware
[14] Chen Jiancheng et al.	Relieves longing by restoring appearance	High real-time requirements	User satisfaction: High	Easy to implement and deploy
[15] Wu et al.	First Chinese multimodal database	Difficult to build and maintain the database	Database size: 48 hours of video and annotations	No need for large-scale database
[16] Zhou et al.	Optimized depression detection with multi-granularity fusion network	Limited effectiveness outside of depression detection	Detection accuracy: Optimized	Wide applicability across fields

PaddleSpeech not only has excellent computing performance and speed, but also has good flexibility and scalability, making it easy to adapt to various application scenarios. Meanwhile, PaddleSpeech adopts the latest deep learning algorithm optimization when processing speech data, which can maintain high recognition accuracy in noisy environments. These advantages make PaddleSpeech the preferred model for achieving efficient key data recognition in this study. By comparing with BigSSL or Wav2Vec, although these models also have their own characteristics, PaddleSpeech shows higher adaptability and effectiveness in terms of application requirements and performance metrics specific to this article. Based on studies conducted by both domestic and international researchers, there is a problem of low detection accuracy in current speech key information recognition technology. Therefore, this research is being conducted to optimize the PaddleSpeech model using speech enhancement algorithms and keyword detection algorithms to improve the recognition performance of speech key information.

3 Construction of speech recognition system and S-PaddleSpeech model

3.1 Construction of PaddleSpeech speech recognition system

The precision of speech key information recognition technology depends on whether the system recognizes speech accurately, and the speech recognition system mainly includes two parts: speech signal (SS) pretreatment and speech recognition. Among them, SS pretreatment is the process of extracting the essence of speech, which includes time-domain waveforms that can reflect many features, but other functional tools are still needed to assist in obtaining more comprehensive information. At present, due to the close fit between Mel-Frequency Cepstral Coefficients (MFCC) and human ear nonlinear frequency induction, MFCC is the most commonly used coefficient in keyword detection technology [19]. When the target speech reaches the frequency band above 800Hz, the speech will attenuate due to the high frequency band, which will result in the energy of the high-frequency SS being less than that of the low-frequency part.

To balance the high and low frequency bands of the SS, it is necessary to emphasize the high-frequency part. The expression of the high pass digital filter for emphasizing the frequency band is shown in equation (1).

$$H(z) = 1 - az^{-1} \quad (1)$$

In equation (1), a represents the pre-emphasis coefficient, and 0.9<a<1.0. To achieve fast Fourier transform, the signal input must be stable. However, due to the time-varying and short-term stationarity of SSs, in order to obtain shorter speech frame segments, it is necessary to perform frame segmentation on the speech. When processing frames, there should be overlap between two consecutive frame segments, that is, frame shift, and the signal segment range should be between 10-30ms. After frame processing, in order to avoid spectral leakage caused by uneven edges of speech frame segments, each speech frame needs to be windowed. Windowing is the process of performing operations on speech frames, resulting in the final frame signal being weakened to 0. Taking the Hamming window as an example, the calculation process of multiplying the window function by the speech frame segment is shown in equation (2).

$$w(n) = 0.54 - 0.46\cos[\frac{2\pi n}{(N-1)}], 0 \le n \le N-1 \quad (2)$$

In equation (2), $w(n)$ represents the window function value of the Hamming window; N represents the window length; n represents the sample index within the window. After windowing, in order to achieve fast Fourier transform, it is necessary to calculate the discrete Fourier transform through a computer. The discrete Fourier transform converts time-based data into data represented in the frequency domain, with sampling points corresponding to complex numbers. The speech frequency-domain signal is the imaginary and real parts of the complex numbers. The specific calculation process of discrete Fourier is shown in equation (3).

$$X(k) = \sum_{N=0}^{N-1} x(n) e^{-j\frac{2\pi nk}{N}}, k = 0,1,...,N-1 \quad (3)$$

In equation (3), the discrete Fourier transform divides the signal into two parts, with the signal length divided from N to N/2. Continuing this process can yield the final result. After fast Fourier transform, the speech is converted from actual frequency to Mel frequency, and then Mel filter bank is used. The correspondence between *Mel* frequency f_{mel} and the original frequency f is shown in equation (4).

$$f_{mei}(f) = 2595\square\lg(1 + \frac{f}{700Hz}) \quad (4)$$

In equation (4), the Mel filter bank primarily consists of M triangular filters, as represented in Figure 1.

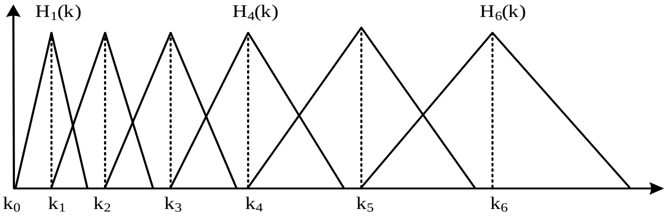


Figure 1: Mel filter bank

In Figure 1, the triangular filters H1 (k), H4 (k), H6 (k), etc. constitute the Mel filter group, which has the characteristics of high frequency with low density and low frequency with high density. This is also in line with the low resolution of the human ear at high frequencies and high resolution at low frequencies. The calculation process of the filter center frequency $f(m)$ is shown in equation (5).

$$f(m) = \frac{N}{f_s} f_{mel}^{-1}\left(f_{mel}(f_l) + m\frac{f_{mel}(f_h) - f_{mel}(f_l)}{M+1} \right), 0 \le m \le M$$

(5)

In equation (5), $f(m)$ represents the frequency dependent function after processing; f_s represents the sampling frequency; f_h represents the highest frequency, and f_l represents the lowest frequency; M represents the length of a processing area or the number of sample points. The calculation method for the coefficient $H_m(k)$ of the filter is shown in equation (5).

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, f(m-1) \le k \le f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, f(m) \le k \le f(m+1) \\ 0, k > f(m+1) \end{cases} \quad (6)$$

In equation (6), $k = 0,1,...,N-1$. After passing through the Mel filter bank, the calculation process of spectral energy $S_i(m)$ is represented in equation (7).

$$S_i(m) = \sum_{k=0}^{N-1} [X_i(k)]^2 H_m(k), 0 \le m \le M \quad (7)$$

After outputting the Mel power spectrum, the calculation of the MFCC is shown in equation (8).

$$mfcc(i,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)\cos\left[\frac{\pi n(2m-1)}{2M}\right] \quad (8)$$

Early speech recognition technology was limited, and its recognition accuracy needed to be improved. However, modern speech recognition systems have made significant progress compared to before, not only increasing the number of speakers recognized, but also greatly enriching the vocabulary [20]. The PaddleSpeech model of Baidu PaddlePaddle series is used for voice identification, and its basic structure is represented in Figure 2.

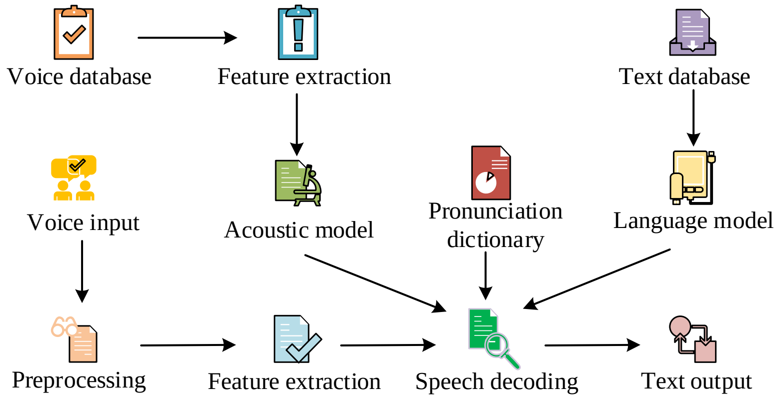


Figure 2: Basic structure of PaddleSpeech model

In the model structure shown in Figure 2, when the speaker speaks, the speech can be converted into electrical signals through a microphone, then converted into data through an analog-to-digital converter, and finally converted from audio to text through the corresponding model.

2.2 Improvement of PaddleSpeech model based on spectral subtraction and wavelet threshold denoising method

In daily environments, SSs are inevitably mixed with noise during input, and the basic PaddleSpeech model is affected by it during speech recognition, thereby reducing the precision of voice identification. Therefore, speech enhancement algorithms improve model precision. Denoising plays a crucial role in keyword recognition, significantly improving the accuracy and efficiency of recognition. Background noise can interfere with the clarity of signals, making it difficult for the system to accurately extract key information when processing audio or video inputs. By denoising, these interferences can be eliminated or reduced, improving the signal-to-noise ratio of the signal and making the keywords more prominent, facilitating the subsequent recognition process.

Speech enhancement algorithms include spectral subtraction and wavelet thresholding denoising. In traditional spectral subtraction, the noisy signal is represented in equation (9).

$$y(t) = x(t) + d(t) \tag{9}$$

In equation (9), $y(t)$ is a noisy signal, $x(t)$ is a pure signal, and $d(t)$ is additive noise. By performing Fourier transform on the three, equation (10) can be obtained.

$$y(\omega) = x(\omega) + d(\omega) \tag{10}$$

By squaring equation (9), the energy distribution of the noisy signal across frequency ranges is shown in equation (11).

$$|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2 + X^*(\omega)D(\omega) + X(\omega)D^*(\omega) \tag{11}$$

In equation (11), $X^*(\omega)D(\omega) + X(\omega)D^*(\omega)$ represents the components produced by the correlation between $X(\omega)$ and $D(\omega)$. Since the prerequisite for spectral subtraction is that $x(t)$ and $d(t)$ are unrelated to each other, equation (12) can be obtained.

$$|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2 \tag{12}$$

Assuming that the noise power of the quiet segment and the speech segment are fixed values, and the estimated noise power of the quiet segment is $|D(\omega)|^2$, the pure speech power spectrum is shown in equation (13).

$$|X(\omega)|^2 = |Y(\omega)|^2 + |D(\omega)|^2 \tag{13}$$

In equation (13), $|X(\omega)|^2 \geq 0$. If it is less than 0, there is a bias in the noise estimation, and the adjustment mechanism is equation (14).

$$\begin{cases} |X(\omega)|^2 = |Y(\omega)|^2 - |D(\omega)|^2, |Y(\omega)|^2 > |D(\omega)|^2 \\ |X(\omega)|^2 = 0, |Y(\omega)|^2 \leq |D(\omega)|^2 \end{cases} \tag{14}$$

After speech enhancement, the approximated frequency distribution of the pure SS is shown in equation (15).

$$\hat{X}(\omega) = |X(\omega)| \exp[j\theta_{Y(\omega)}] \tag{15}$$

In equation (15), $\theta_{Y(\omega)}$ represents the phase of noisy speech, and $\hat{X}(\omega)$ represents the SS without noise. Performing inverse Fourier transform on $\hat{X}(\omega)$ results in an enhanced SS as shown in equation (16).

$$\hat{x}(\omega) = IFFT\left[\hat{X}(\omega)\right] \tag{16}$$

Spectral subtraction has a small computational complexity and is widely used in speech enhancement. However, there is also a problem that SSs may contain musical noise. To reduce its impact, the adjustment mechanism is improved, as shown in equation (17).

$$\begin{cases} |X(\omega)|^2 = |Y(\omega)|^2 - \alpha |D(\omega)|^2, |Y(\omega)|^2 > \alpha |D(\omega)|^2 \\ |X(\omega)|^2 = \beta |D(\omega)|^2, |Y(\omega)|^2 \leq \alpha |D(\omega)|^2 \end{cases} \tag{17}$$

In equation (17), α is the over reduction factor used to control the degree of noise reduction. Increasing α helps improve the signal-to-noise ratio and suppress noise. β is a compensation factor that determines the amount of noise retained in the spectrum. Increasing β will reduce the noise component, but too much may lead to a rise in ambient noise levels. The different values of α and β affect the denoising effect. In traditional methods, α=1 and β=0. Research has optimized traditional spectral subtraction by adjusting these two parameters, namely α=5 and β=0.002. Through preliminary experiments, it was found that the excessive reduction factor value can effectively suppress noise without significantly distorting the speech signal. This compensation factor value achieves a balance between noise suppression and background sound preservation, avoiding the environmental noise increase caused by excessive suppression. This choice is universal because its moderate residual noise allows for the reduction of background noise interference while preserving important speech features. In practical scenarios, such as phone conversations or meeting minutes, this parameter can balance speech intelligibility with background realism. In the key process of spectrum subtraction adjustment, in addition to optimizing parameters, the characteristics of environmental noise and the spectral features of signals are also crucial. Firstly, experiments need to be conducted under different signal-to-noise ratio conditions to evaluate the adaptability of adjusting parameters to different types of noise. This means that in practical applications, it is necessary to establish a dynamic adjustment mechanism based on specific scenarios, so that parameter settings can be automatically adjusted according to external noise changes in real-time processing.

The theory of wavelet transform can localize the time-frequency of signals and has the characteristic of flexibly selecting wavelet functions. Wavelet threshold denoising aims to enhance valuable components in SSs. In wavelet coefficients, useful signals are usually composed of larger amplitude components [21]. The key to achieving wavelet threshold denoising lies in selecting appropriate wavelet bases, thresholds, and threshold functions. The selected wavelet in the study is symN, which may cause phase distortion when reconstructing SSs. However, symN wavelet can reduce phase distortion to a certain extent. Regarding threshold selection, the threshold chosen for the study is the minimax threshold, which is calculated as shown in equation (18).

$$\lambda\begin{cases} 0, N \le 32 \\ 0.3936 + 0.1829\dfrac{\ln N}{\ln 2}, N > 32 \end{cases} \tag{18}$$

The threshold functions are generally categorized into two types: hard and soft, which are commonly used to correct wavelet coefficient errors. The hard threshold image is shown in Figure 3.

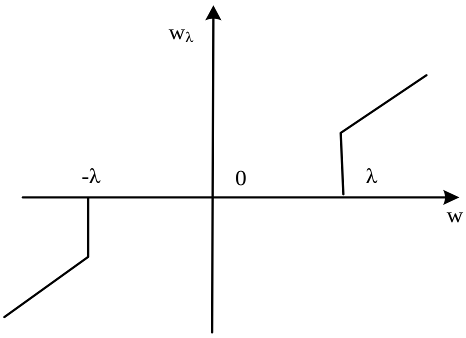


Figure 3: Hard threshold function image

The soft threshold image is shown in Figure 4.

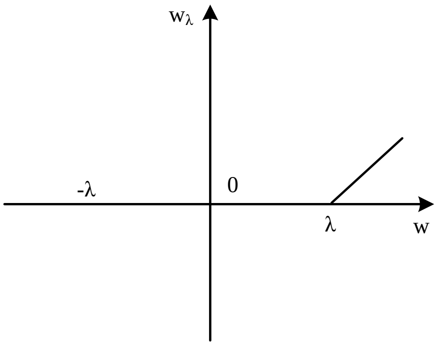


Figure 4: Soft threshold function image

The threshold function expression selected by the research is shown in equation (19).

$$\hat{w}_j = \begin{cases} w_j, \left|w_j\right| > \beta\lambda \\ \operatorname{sgn}(w_j)(\left|w_j\right| - \lambda)\dfrac{\beta}{\beta - 1}, \lambda < \left|w_j\right| < \beta\lambda \\ 0, \left|w_j\right| \le \lambda \end{cases} \tag{19}$$

In equation (19), $\hat{w}_j$ represents the estimated value of wavelet coefficients, j represents the decomposition scale, w_j represents the original wavelet coefficients, β represents a constant, and $1.5 \le \beta \le 3$. λ represents the threshold. Research enhances speech through spectral subtraction and denoises speech through wavelet transform, in order to optimize the initial PaddleSpeech model and improve its speech recognition accuracy, ultimately obtaining the S-PaddleSpeech model. The main differences in technology and framework between PaddleSpeech and S-PaddleSpeech are reflected in the following aspects: Firstly, S-PaddleSpeech adopts a Transformer based model architecture, while PaddleSpeech uses traditional CNN and RNN structures. Secondly, S-PaddleSpeech implements optimization strategies such as mixed precision training and distributed training to improve training efficiency, while PaddleSpeech relies on relatively fixed training methods.

In terms of feature extraction, S-PaddleSpeech emphasizes end-to-end learning, reducing reliance on traditional feature extraction techniques. In addition, S-PaddleSpeech has a higher modular design that allows users to flexibly choose and replace model components, while PaddleSpeech is relatively lacking in this regard. Finally, S-PaddleSpeech supports advanced features such as multi speaker recognition and emotion recognition, while PaddleSpeech mainly focuses on basic speech recognition and synthesis tasks. These core technological differences have built a more flexible and efficient framework for S-PaddleSpeech.

Speech recognition technology can convert speech as a whole into text, but to recognize key information, keyword detection algorithms need to be introduced to accurately extract the required information. Therefore, the research will combine deep learning-based keyword recognition algorithms with the S-PaddleSpeech model to further optimize it. The process of speech keyword detection technology is represented in Figure 5.

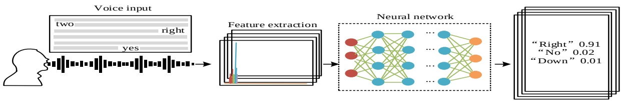


Figure 5: Process of speech keyword detection technology

In Figure 5, the speech keyword detection technology first extracts characteristics from the input SS, and then trains the neural network model. After that, the neural network model extracts vectors from the features and classifies them. Finally, the probability distribution of speech is analyzed, and the keyword with the highest probability is the required result. Due to the strong feature extraction ability and small number of training parameters of CNN, this study uses CNN to optimize the S-PaddleSpeech model. Its structure is shown in Figure 6 [22].

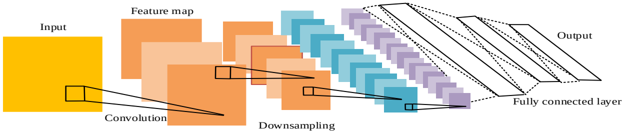


Figure 6: CNN structure diagram

In Figure 6, the hidden layers of CNN models are generally divided into convolutional layers and pooling layers. The convolutional layer is the core part of CNN, which performs dimensionality reduction and feature extraction on the input image through convolution operations (linear operations, i.e., shifting on the original image). The function of the pooling layer is feature extraction, which involves removing certain features. By eliminating non-essential samples from the feature map, the parameter count can be decreased further. After the SS processing is completed, the CNN model extracts feature from the MFCC matrix through convolutional kernels, and then reduces the size of the feature map through pooling layer aggregation operations. Finally, the feature vectors are transformed into one dimension and the final result is output to obtain the C-S-PaddleSpeech model. Before applying the algorithm in practice, there is also a data preprocessing step, which is mainly divided into three parts: frame segmentation and windowing, Fourier transform and Mel filtering, and noise level estimation. Divide the speech signal into frames to ensure short-term stability, with a frame length of 25 milliseconds and a frame shift of 10 milliseconds. Each frame is windowed using Hamming window to reduce spectral leakage in subsequent Fourier analysis. Perform discrete Fourier transform on the windowed signal of each frame to extract frequency domain features.

3 Application testing of C-S-PaddleSpeech model

3.1 Performance testing of speech recognition algorithms and keyword algorithms

To verify the capability of the speech enhancement algorithm, the ROC curve was used for performance testing. The experimental system used Windows 11, the device was a 64-bit operating system, and the programming language was Python. The display processor of the experimental equipment is NVIDIA GeForce RTX 3080, the processor is Intel Core i7-11700K, the memory is 32GB DDR4 3200MHz RAM, and the storage device is 1TB NVMe SSD. The dataset used in the study includes speech samples from different fields and environments, such as indoor conversations, street noise, coffee shop background noise, and industrial area noise. The speech types cover speakers of various genders, ages, and accents, ensuring sample diversity. The dataset consists of 5000 recording samples, ranging in duration from 3 seconds to 30 seconds, with a total duration of approximately 30 hours. The noise level has been carefully labeled and divided into low (5-15 dB SNR), medium (0-5 dB SNR), and high (below 0 dB SNR) noise environments to ensure the presence of diverse noise influences in the dataset. The noise level has been carefully labeled and divided into low (5-15 dB SNR), medium (0-5 dB SNR), and high (below 0 dB SNR) noise environments to ensure the presence of diverse noise influences in the dataset. The use of 5-fold cross validation effectively reduces the risk of model overfitting caused by random partitioning, ensuring the algorithm's generalization ability on unseen data. In order to promote the reproducibility of the experiment, the random seed used in the experiment was 123456. Fixed random seeds ensure that the random number sequence generated during each experiment run is the same, allowing different researchers to replicate the experimental results under the same conditions. The deep learning framework applied in the experiment was TensorFlow. The test results are shown in Figure 7.

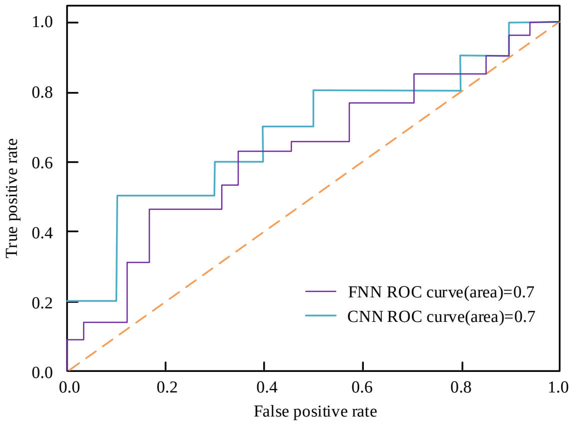


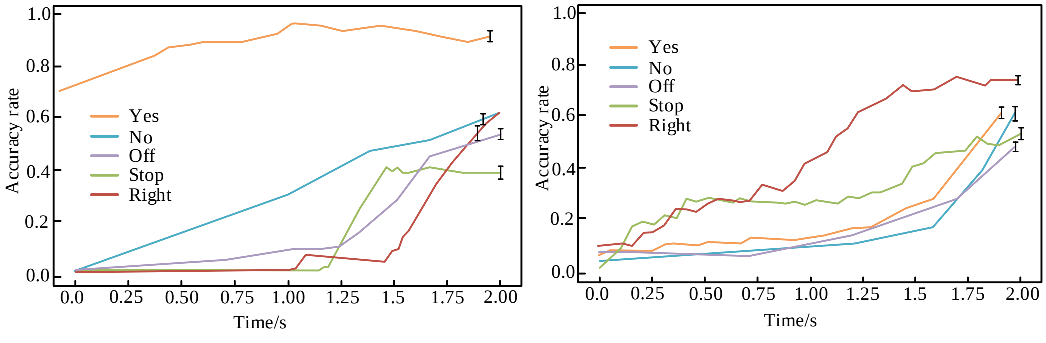
Figure 7: ROC schematic diagram

Figure 7 shows the False Positive Rate (FPR) as the x-axis and the True Positive Rate (TPR) as the y-axis. The closer the ROC curve is to the upper left corner, the better the performance of the algorithm. Observing Figure 7, it can be seen that the ROC curve of CNN shows an overall stepwise upward trend, especially in the early stage where the curve rises rapidly, and the upward speed slows down slightly in the middle and later stages, but still maintains a stable upward trend. In contrast, the FNN curve is closer to the bottom right corner. These ROC curves can clearly depict the classification performance of the model at different thresholds, which is consistent with the theoretical prediction of the relationship between signal-to-noise ratio and model performance, that is, a higher signal-to-noise ratio will significantly reduce the interference of noise on the model, thereby improving recognition rate. Through t-test on the result data, it was found that the p-values of the obtained test data were all below the significance level of 0.05. From this, the speech enhancement algorithm raised by the study could effectively enhance speech and had good noise reduction effect. To confirm the performance of the keyword detection algorithm, the same keywords in long and short sentences were individually evaluated, and the outcomes are represented in Figure 8.

In Figure 8 (a), in the initial stage, the algorithm had a higher detection accuracy for the keyword “yes” in short sentences compared to other keywords. Its accuracy curve fluctuated up and down over time, but the accuracy remained above 0.7. The detection accuracy of keywords “stop” and “right” was relatively low in the first 1 second, and the accuracy curve showed a rapid growth trend after 1 second. This reflects the algorithm's strong ability to detect specific keywords in a short period of time. In Figure 8 (b), the accuracy curves of the keywords "right" and "stop" in long sentences fluctuate up and down. Although its overall accuracy is gradually improving, it is generally lower than the accuracy in short sentences. This trend may be related to the complexity and information content of sentences, as long sentences often contain more information, leading to certain interference in algorithm recognition. The accuracy of keyword detection is above 0.5 for both long and short sentences. Through t-test on the result data, it was found that the p-values of the obtained test data were all below the significance level of 0.05. The CNN keyword algorithm proposed in this study has good performance in detecting keywords in speech.

3.2 Comparative testing of speech detection algorithms and keyword detection algorithms

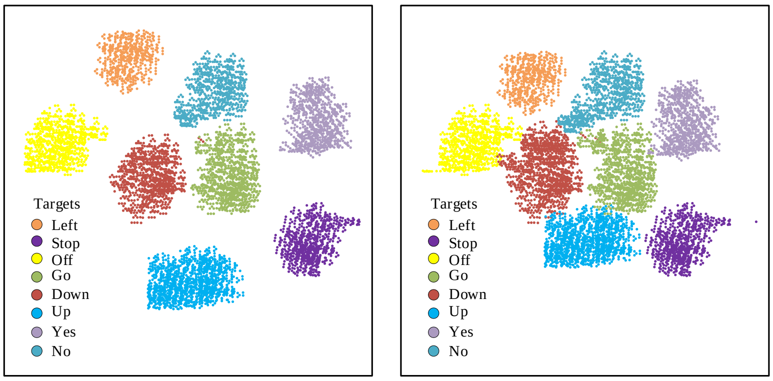
To further verify the effectiveness of speech detection algorithms and keyword detection algorithms in improving PaddleSpeech model speech detection, the study compared them with other algorithms separately. Firstly, the speech enhancement algorithm based on spectral subtraction and wavelet threshold denoising proposed by the research was compared with the Kalman filter method. The experimental data was obtained from a large conference speech text, and the results are shown in Figure 9.



(a) Short sentence keyword detection

(b) Long sentence keyword detection

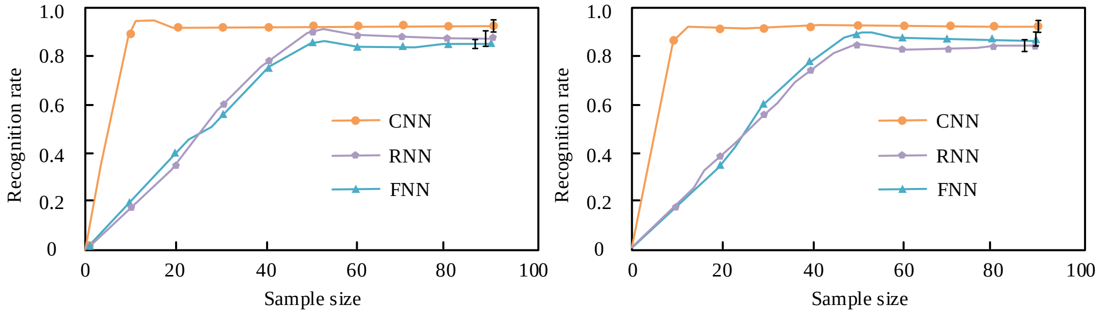
Figure 8: Results of keyword detection for long and short sentences



(a) Speech enhancement algorithm proposed by the research institute

(b) Kalman filter algorithm

Figure 9: Comparison results of two algorithms



(a) Specific person

(b) Non-specific person

Figure 10: The influence of sample size on recognition rate

As shown in Figure 9 (a), there were significant differences in the distance between clusters of different speech texts. Specifically, the distance between clusters “go” and “down” was relatively close, while the distance between clusters “go” and “up” was relatively far. This indicates that speech enhancement algorithms reduced the distance between similar speech text categories and increased the distance between speech texts with significant differences by enhancing speech and denoising methods. In Figure 9 (b), the cluster center representing “off” was closer to the cluster center representing “yes”, and there was a significant cross fusion phenomenon between the two clusters. This indicates that the Kalman filter method had lower recognition accuracy than the proposed speech enhancement algorithm in distinguishing similar speech text categories. Overall, the feature space distribution in Figure 9 (a) was more “ideal”, with clearer boundaries between different speech texts, which was more conducive to improving the model's recognition ability. For keyword detection algorithms, the fewer samples required while ensuring recognition rate, the better. The accuracy of keyword detection is above 0.5 for both long and short sentences. Through t-test on the result data, it was found that the p-values of the obtained test data were all below the significance level of 0.05. Therefore, this study took specific person speech and non-specific person speech as examples, and used CNN, RNN, and FNN algorithms for keyword detection. The results are shown in Figure 10.

In Figure 10 (a), for a specific person's speech, the CNN algorithm exhibited a high recognition rate when the sample size was less than 20 in the early stage, and its recognition rate curve showed a rapid growth trend, with a stable recognition rate of 0.9 in the later stage. In contrast, the recognition rate curves of RNN and FNN algorithms grew slightly slower, and their recognition rates only stabilized at 0.8 when the sample size reached 60. Both algorithms had a larger sample demand than CNN algorithms and lower recognition rates. Figure 10 (b) shows the recognition rate curve of the three algorithms under non-specific human speech conditions. The case of non-specific speech, the impact of sample size and noise on the algorithm was not much different from that of specific people, and its recognition rate was not significantly reduced compared to specific person speech. The accuracy of keyword detection is above 0.5 for both long and short sentences. Through t-test on the result data, it was found that the p-values of the obtained test data were all below the significance level of 0.05. From this, the CNN algorithm was equally applicable to both specific and non-specific speech. The main reason for choosing CNN over RNN and FNN is that CNN has higher computational efficiency and better performance when processing large-scale datasets, especially suitable for complex tasks such as image and natural language processing. Secondly, CNN exhibits stronger adaptability, capable of automatically learning data features and flexibly adapting to various application scenarios. In addition, CNN can better utilize the parallel processing capabilities of modern hardware such as GPUs, thereby improving training and inference efficiency. Finally, CNN typically provides more stable and reliable results in multiple experiments. These advantages make CNN a better choice in practical applications.

3.3 Application testing of C-S AddleSpeech model

To test the validity of the C-S AddleSpeech model in speech keyword detection, the study first set a total of 20 keywords in multiple audios and compared the C-S AddleSpeech model before and after improvement. The results are shown in Figure 11.

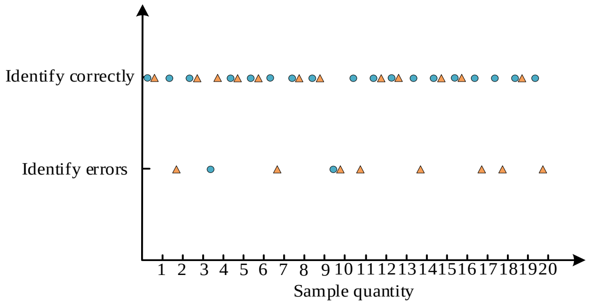


Figure 11: Comparison of recognition results between two models

In Figure 11, for the 20 keywords in the audio, the PaddleSpeech model recognized 12 correctly, while the C-S-PaddleSpeech model recognized 18 correctly. The former had an accuracy rate of 60%, while the latter had an accuracy rate of 90%, with a difference of 30% between the two. From this, the PaddleSpeech model significantly improved its ability to recognize keywords after keyword recognition optimization. The accuracy of keyword detection is above 0.5 for both long and short sentences. Through t-test on the result data, it was found that the p-values of the obtained test data were all below the significance level of 0.05. To test the recognition performance of the C-S AddleSpeech model in speech and audio, a comparison test was performed between the C-S AddleSpeech model and the SpeechRecognition model proposed by Google. The results are shown in Table 2.

Table 2: Comparison of C-S-PaddleSpeech and SpeechRecognition

Model	Total word count	Missing number	The number of multiple words	Error word count	Correct word count	Word accuracy rate (%)
C-S-PaddleSpeech model	15177	25	2	158	15019	98.9589
SpeechRecognition model	15177	54	9	584	14593	96.1520

In Table 2, the speech audio contained a total of 15177 characters. The C-S-PaddleSpeech model correctly recognized 15019 characters, with 25 missing words and 158 incorrect words. The overall recognition rate was high, reaching 98.9580%. The SpeechRecognition model correctly recognized 14593 characters, with 54 missing words and 584 incorrect words. The accuracy of keyword detection is above 0.5 for both long and short sentences. Through t-test on the result data, it was found that the p-values of the obtained test data were all below the significance level of 0.05. Both missing and incorrect words were at least twice as many as the C-S AddleSpeech model. From this, after optimizing the speech enhancement and keyword detection algorithms, the C-S-PaddleSpeech model had a high accuracy and excellent performance in speech recognition. In order to further determine the superiority of the research method, performance tests were conducted on the LibriSpeech dataset, as shown in Table 3.

Table 3: LibriSpeech dataset testing

Metric	CNN (research method)	RNN	Transformers
Dataset	LibriSpeech	LibriSpeech	LibriSpeech
Accuracy (%)	92.5	89	87.5
Training Time (hours)	5	7	6
Testing Time (ms/sample)	25	40	35
Memory Usage (GB)	3.2	4.5	4.0
Model Complexity (Million Parameters)	15	20	18

From Table 3, it can be seen that CNN outperforms RNN and Transformers methods in terms of accuracy, computation time, and memory usage. On the LibriSpeech dataset, the accuracy of CNN is 92.5%, significantly higher than RNN's 89.0% and Transformers's 87.5%. Through t-test on accuracy data, the results showed that the p-value between CNN and RNN was 0.002, and the p-value between CNN and FNN was 0.001, both of which were less than the significance level of 0.05, indicating that CNN was significantly better than the other two methods in accuracy. In addition, the training and testing time of CNN are superior to RNN and Transformers indicating that it has a greater advantage in computational efficiency. Especially on large-scale datasets, the memory footprint of the model is smaller, making the research method more suitable for practical applications.

3.3 Discussion

The PaddleSpeech model has been improved by combining spectral subtraction and wavelet threshold denoising algorithms, significantly enhancing the accuracy of speech key information recognition. Compared with the LSTM algorithm used in related work (with a recognition accuracy of 94%), the research method demonstrates impressive performance in processing specific and non-specific speech. When the sample size is less than 20, the research method achieves a recognition accuracy of 90%. This indicates that the research method is superior to previous methods in the field of Chinese speech recognition. The technology studied has shown excellent performance in dealing with noise interference. By applying spectral subtraction, the signal-to-noise ratio can be effectively improved, making the extraction of key information clearer and reducing the negative impact of background noise on speech recognition. In addition, compared with the traditional Kalman filtering method, it was found that the acoustic feature space used was more ideal, effectively improving the discrimination of similar speech categories. Although research methods have advantages in accuracy and efficiency, their limitations cannot be ignored. For example, recognition accuracy may be affected in multiple languages or complex speech environments. This issue is particularly evident in multilingual conversations and may lead to a decrease in model performance. Therefore, future research can shift the focus to language adaptation optimization and introduce dynamic adjustment mechanisms to enhance application capabilities in various environments. In addition, multimodal fusion techniques can be explored, such as combining video information with speech data, to further enhance the accuracy and user experience of recognition systems. In summary, the study provides an effective solution for speech key information recognition technology by improving the PaddleSpeech model, promoting the development of this field and laying the foundation for future research directions.

4 Conclusion

In addressing the issue of insufficient accuracy in speech key information detection technology, a PaddleSpeech model was developed and optimized by integrating speech enhancement technology and keyword detection technology to raise the identification accuracy of speech key information. The experiment outcomes showed that for the same keywords in long and short sentences, the accuracy of keyword detection technology exceeded 0.5, demonstrating good detection performance. In addition, in the detection of specific and non-specific human speech, the CNN detection algorithm achieved a keyword recognition accuracy of 0.9 when the sample size was less than 20, while the FNN algorithm achieved an accuracy of 0.8 when the sample size reached 60. This indicates that CNN is superior to FNN in both sample size requirements and recognition accuracy. In the application testing of the model, the improved PaddleSpeech model performed significantly better than before in identifying 20 keywords in audio, with an accuracy rate of up to 90%. In comparison with the SpeechRecognition model for audio character recognition, the improved PaddleSpeech model correctly recognized 15019 characters with an accuracy of 98.9580%, while the SpeechRecognition model correctly recognized 14593 characters with an accuracy of 96.1520%, which was 2.806% higher than the latter. From this, the improved PaddleSpeech model raised in the study could effectively improve the accuracy of speech key information recognition. However, there are also some shortcomings in the research, such as the decrease in detection accuracy of the model due to differences in language rules or language structures in multilingual conversation scenarios, which need to be improved in future research.

References

[1] Žvirblis T, Pikšrys A, Bzinkowski D, Rucki M, Kilikevič ius A, Kurasova O (2024). Data augmentation for classification of multi-domain tension signals. Informatica, pp. 883-908. https://doi.org/10.15388/24-INFOR578

[2] Adolfi F, Bowers J S, Poeppel D (2023). Successes and critical failures of neural networks in capturing human-like speech recognition. Neural Networks, pp. 199-211. https://doi.org/10.1016/j.neunet.2023.02.032

[3] Chen, Y. (2024). Human Resource Recommendation Based on CBCF-BAC and Short Text Similarity Algorithm. Informatica, pp. 99-112. https://doi.org/10.31449/inf.v48i22.6853

[4] Mena-Yedra R, López Redondo J, Pérez-Sánchez H, Martinez Ortigosa P (2024). ALMERIA: Boosting pairwise molecular contrasts with scalable methods. Informatica, pp. 617-648. https://doi.org/10.15388/24-INFOR558

[5] Navakauskas D, Kazlauskas M (2023). Fog computing in healthcare: systematic review. Informatica, pp. 577-602. https://doi.org/10.15388/23-INFOR525

[6] Osipov A, Pleshakova E, Liu Y (2024). Machine learning methods for speech emotion recognition on telecommunication systems. Journal of Computer Virology and Hacking Techniques, pp. 415-428. https://doi.org/10.1007/s11416-023-00500-2

[7] Kumar P, Malik S, Raman B (2024). Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. Multimedia Tools and Applications, pp. 28373-28394. https://doi.org/10.1007/s11042-023-16443-1

[8] Belkebir, M., Maarouk, T. M., & Nini, B. (2024).

Real-time Semantic Healthcare System: Visual Risks Identification for Elders and Children. Informatica, pp. 65-82. https://doi.org/10.31449/inf.v48i14.6271

[9] Yuan Q, Dai Y, Li G (2023). Exploration of English speech translation recognition based on the LSTM RNN algorithm. Neural Computing and Applications, pp. 24961-24970. https://doi.org/10.1007/s00521-023-08462-8

[10] Xu Y, Su H, Ma G, Liu X (2023). A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context. Complex & Intelligent Systems, pp. 951-963. https://doi.org/10.1007/s40747-022-00841-3

[11] Zhang Y, Park D S, Han W (2022). Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. IEEE Journal of Selected Topics in Signal Processing, pp. 1519-1532. https://doi.org/10.1109/JSTSP.2022.3182537

[12] Zhu P, Cheng D, Yang F (2022). Improving Chinese named entity recognition by large-scale syntactic dependency graph. IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 979-991. https://doi.org/10.1109/TASLP.2022.3153261

[13] Ambrogio S, Narayanan P, Okazaki A (2023). An analog-AI chip for energy-efficient speech recognition and transcription. Nature, pp. 768-775. https://doi.org/10.1038/s41586-023-06337-5

[14] Burchi M, Timofte R (2023). Audio-visual efficient conformer for robust speech recognition. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2258-2267. https://doi.org/10.1109/WACV56688.2023.00229

[15] Deng K, Gao D, Guan Q, Wang F, Ma S, Zhao C, Lu Y (2023). Exploring the effectiveness of using a smartphone for position-oriented process monitoring. The International Journal of Advanced Manufacturing Technology, pp. 4293-4307. https://doi.org/10.1007/s00170-023-10984-3

[16] Chen J C, Liao C Y, Chen M H (2024). Using Virtual Reality Technologies to Cope with Grief and Emotional Resilience of Absent Loved Ones. Proceedings of the 2024 8th International Conference on Medical and Health Informatics, pp. 73-77. https://doi.org/10.1145/3673971.3674018

[17] Wu J, Chen S, Xiang W, Sun L, Zhang H, Zhang Z, Li Y (2024). CNAMD Corpus: A Chinese Natural Audiovisual Multimodal Database of Conversations for Social Interactive Agents. International Journal of Human－Computer Interaction, pp. 2041-2053. https://doi.org/10.1080/10447318.2023.2228530

[18] Zhou L, Liu Z, Li Y (2024). Multi Fine-Grained Fusion Network for Depression Detection. ACM Transactions on Multimedia Computing, Communications and Applications, pp. 1-23. https://doi.org/10.1145/3665247

[19] Atmaja B T, Sasou A, Akagi M (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. Speech Communication, pp. 11-28. https://doi.org/10.1016/j.specom.2022.03.002

[20] Melnik-Leroy G A, Bernatavičienė J, Korvel G, Navickas G, Tamulevičius G, Treigys P (2022). An overview of Lithuanian intonation: a linguistic and modelling perspective. Informatica, pp. 795-832. https://doi.org/10.15388/22-INFOR502

[21] Ok M W, Rao K, Pennington J (2022). Speech recognition technology for writing: usage patterns and perceptions of students with high incidence disabilities. Journal of Special Education Technology, pp. 191-202. https://doi.org/10.1177/0162643420979929

[22] Blanco-Fernńndez Y, Gil-Solla A, Pazos-Arias J J, Quisi-Peralta D (2023). Automatically Assembling a Custom-Built Training Corpus for Improving the Learning of In-Domain Word/Document Embeddings. Informatica, pp. 491-527. https://doi.org/10.15388/23-INFOR527