Research and Implementation of Low Resource Voice Awakening Technology in Smart Home Scene

Zhuoxi Li[*]

School of Computing, Guangdong Neusoft University, Foshan, China
15122597559@163.com
[*]*Corresponding author*

Abstract: *With the rapid development of artificial intelligence and Internet of Things technology, smart home has become an important part of modern life. The smart home system greatly improves the convenience and comfort of users through voice control, intelligent control and other technologies. However, in practical applications, smart home devices face resource constraints, especially in terms of battery life and computing power. Voice wake-up technology in low resource environment is of great significance to smart home system. Low-power voice wake-up technology can extend the battery life of smart home devices and improve the user experience. Secondly, the voice wake-up technology with low computational complexity can reduce the hardware cost of smart home devices and promote the popularity of smart home systems. Efficient and accurate voice wake-up technology can also improve the interaction efficiency and user satisfaction of smart home systems. Therefore, the study of voice wake-up technology in low resource environment is of great significance to promote the development and application of smart home system. At present, voice awakening technology in low resource environment has become a research hotspot in academia and industry. Domestic and foreign scholars have done a lot of research on feature extraction, model optimization, low power design and so on. In feature extraction, Mel-Frequency Cepstral Coefficients (MFCC) and other algorithms are widely used in speech signal processing. In terms of model optimization, Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) and other models are used to improve the accuracy and robustness of voice arousal. In terms of low power design, scholars have proposed methods such as event-driven wake-up mechanism and low power hardware accelerator to reduce the power consumption of voice wake-up systems. However, there are still some problems with the existing research. The existing feature extraction algorithms and model optimization methods are not effective in high noise environment, which can easily lead to false wake up and missed wake up. Secondly, low power design often requires a trade-off between wake up rate and power consumption, and how to reduce power consumption while ensuring wake up rate is a difficult problem. In addition, most of the existing voice wake-up systems are designed for specific scenarios and lack generality and scalability. In this paper, low resource voice wake-up technology in smart home scene is studied, and a method based on low power feature extraction and neural network optimization is proposed. Aiming at the resource limitation problem of smart home devices, low power and efficient feature extraction algorithms, such as MFCC and its improved algorithm, are studied. The study of neural network model optimization applies to neural network models in low-resource environments, such as deep neural network (DNN), convolutional neural network (CNN), etc., and performs model optimization to improve the accuracy and robustness of voice arousal. Low power design methods such as low power hardware accelerators and event-driven wake up mechanism are studied to reduce the power consumption of voice wake up systems. The research of this paper provides a new idea and method for voice wake-up technology in smart home system, which is of great significance for promoting the development and application of smart home system.*

Keywords: *Smart Home; Voice Wake-Up; Low Power Consumption; Neural Network; Resource Constraint*

1. Introduction

Through the Internet of Things technology, the smart home system connects various household appliances, security systems, lighting systems, etc., to achieve intelligent management and control. As one of the important interaction ways of smart home, voice control is favored because of its natural and convenient characteristics. However, smart home devices face resource constraints in practical applications, especially in terms of battery life and computing power. How to realize efficient and accurate voice wake-up in low resource environment has become a key issue in the research and implementation of smart home system.

2. Voice Wake-up Technology in Smart Home Scenarios

2.1 Overview of Voice Wake Technology

Voice wake technology is a technology based on keyword recognition, which is used to detect a specific wake word in a continuous speech stream and wake the system from sleep. As one of the important interaction ways of smart home system, voice wake-up technology has the characteristics of natural and convenient [1]. Through voice wake technology, users can control smart home devices through voice commands without touching the device, improving the user experience.

2.2 Principles of the Voice Wake System

The voice wake-up system is mainly composed of speech preprocessing, feature extraction, model matching and result output. Speech preprocessing is to preprocess the input speech signal, including denoising, filtering, endpoint detection, etc., in order to improve the accuracy of subsequent feature extraction and model matching. Feature extraction is to extract feature vectors from pre-processed speech signals. Commonly used feature extraction algorithms include Mayer cepstrum coefficient (MFCC), linear predictive cepstrum coefficient (LPCC), etc [2]. Model matching matches the extracted feature vector with the pre-trained wakeword model to determine whether there is a wakeword. The commonly used model matching algorithms include dynamic time warping (DTW), Hidden Markov Model (HMM), deep neural network (DNN) and so on. The final result output is based on the result of the model matching, output the wake signal or reject signal to control the state of the smart home device.

2.3 Voice Wake-up Technology Challenges in Smart Home Scenarios

In the smart home scenario, voice wake-up technology faces several challenges. First of all, noise interference has become a major obstacle, because smart home devices are often placed in the home environment, vulnerable to the impact of TV sound, talk sound, kitchen noise and other noises, and then interfere with the normal operation of the voice wake-up system, causing false wake up and missing wake up phenomenon. Secondly, the problem of resource limitation is significant, in view of the limitations of smart home devices in computing power and battery life, it has become an urgent problem to achieve low power consumption and low computing complexity while ensuring wake up rate. In addition, the application in the multi-user scenario also brings challenges. How to effectively distinguish the voice commands when the voice wake-up system is used by multiple users at the same time and improve the accuracy and robustness of the system becomes the key. Finally, privacy protection cannot be ignored, given that voice wake systems need to process user voice messages, ensuring user privacy and data security becomes a crucial issue.

3. Research on Low-power Feature Extraction Algorithm

3.1 Mayer Cepstrum Coefficient (MFCC) Algorithm

Mayer cepstrum coefficient (MFCC) is a commonly used speech feature extraction algorithm, which has the advantages of strong robustness to noise and low computational complexity. The MFCC algorithm extracts the key features of speech signals by converting them into cepstrum coefficients on the Meir frequency scale.

(1) Principle of MFCC algorithm

The MFCC algorithm mainly includes the following steps: First, the input speech signal is pre-weighted to improve the energy of the high frequency part and reduce the spectrum distortion. The second is to divide the voice signal into frames with a length of 20-30ms, and window processing is performed on each frame to reduce spectrum leakage. The third is the fast Fourier transform (FFT), which carries on the fast Fourier transform to each frame speech signal to obtain the spectrum information. The fourth is the Meir filter bank, the spectrum information is filtered through the Meir filter bank, and the Meir spectrum is obtained. The fifth is the logarithmic transformation, the logarithmic transformation of the MEL spectrum, to obtain the logarithmic MEL spectrum. The sixth is the discrete cosine transform (DCT), the logarithmic MFCC coefficient is obtained by DCT.

(2) Improvement of MFCC algorithm

To solve the problem of noise interference and resource limitation in the smart home scene, the MFCC algorithm can be optimized. In the pre-processing stage, adaptive filtering, spectral subtraction and other technologies are applied to implement noise suppression of speech signals to enhance the accuracy of feature extraction [3]. After the MFCC coefficients were extracted, principal component analysis (PCA) or linear discriminant analysis (LDA) were used to reduce the dimensionality of the features to reduce the computational complexity. In addition, by calculating the first-order and second-order differences of the MFCC coefficients, the timing features of the speech signal can be effectively captured, thus improving the robustness of the model.

3.2 Low Power Feature Extraction Algorithm

(1) Event-driven feature extraction

Event-driven feature extraction algorithm is an efficient speech signal processing method. The core of this method is to accurately detect key events in speech signals, which usually include the start point, end point of speech and other critical moments for subsequent processing [4]. When these critical events are detected, the algorithm will immediately carry out feature extraction, without the need to process the entire speech signal indiscriminately. This strategy significantly reduces the amount of unnecessary computation, resulting in low power consumption and excellent real-time performance. In smart home devices, the algorithm can efficiently process speech signals and provide strong support for subsequent speech recognition and response.

(2) Low complexity feature extraction

The core of low complexity feature extraction algorithm is to simplify the feature extraction process, thus greatly reducing the required computing resources. For example, in the selection of filter banks, the low complexity feature extraction algorithm may use a simpler filter bank to replace the complex MEL filter bank. Such substitution significantly reduces the amount of computation while maintaining certain accuracy. In addition, the algorithm may also introduce advanced technologies such as approximate computation and quantization to further reduce the computational complexity, so as to meet the strict requirements of computing performance and power consumption of smart home devices.

(3) Feature extraction based on hardware acceleration

The feature extraction algorithm based on hardware acceleration makes full use of the powerful computing power of low power hardware accelerators (such as DSP, FPGA, etc.), and effectively implements the feature extraction algorithm on these hardware platforms. Through hardware acceleration, the speed of feature extraction is significantly improved, and the power consumption is effectively controlled. This method is especially suitable for smart home devices that require high computing performance and power consumption, providing them with powerful and reliable voice signal processing support. In practical applications, the feature extraction algorithm based on hardware acceleration has achieved remarkable results, which provides a strong guarantee for the intelligence and convenience of smart home devices.

4. Research on Neural Network Model Optimization

4.1 Deep Neural Network (DNN) Model

Deep neural network (DNN) is a common neural network model with strong learning ability and generalization ability, which is suitable for complex speech signal processing tasks. In voice arousal technology, DNN model can be used for feature vector classification and recognition to determine whether the input voice signal contains a specific wake-up word [5].

(1) Principle of DNN model

The DNN model consists of multiple layers of neurons, including input layer, hidden layer and output layer. The input layer receives the pre-processed feature vectors, the hidden layer extracts high-level features through nonlinear transformation, and the output layer outputs the classification results [6]. DNN models are trained using a backpropagation algorithm that constantly adjusts the weights between neurons to minimize classification errors.

(2) Application of DNN model in voice wake-up

In voice wake up technology, DNN model can be used to construct wake up word classifier. The process begins with feature extraction of pre-processed speech signals. Feature extraction is a key step in speech signal processing, which can transform the original speech signal into a series of vectors that can reflect the speech characteristics. These feature vectors are then input into the designed DNN model for further classification processing. DNN model can output a probability distribution according to the input eigenvector with its powerful nonlinear mapping capability. This probability distribution visually shows the likelihood that the input speech signal belongs to each preset category. In the wake word classifier scenario, these categories usually correspond to different wake words or background noise, etc. In order to accurately determine whether the input speech signal contains a specific wake word, a reasonable threshold is usually set. When the probability of a certain wake word class output by DNN model exceeds this threshold, it can be considered that the input voice signal contains the wake word, thus triggering the corresponding wake operation. This method not only improves the accuracy of voice wake-up, but also enhances the robustness and practicability of the system.

(3) Optimization of DNN model

DNN model can be optimized to solve the resource limitation problem in smart home scenario. Firstly, pruning and quantization techniques can be used to reduce the number of parameters and calculation amount of the model, and reduce the complexity and power consumption of the model. Secondly, knowledge distillation and other techniques can be used to transfer the knowledge of the large model to the small model to maintain the performance of the model and reduce the computational complexity. In addition, distributed training and model compression techniques can be used to further optimize the performance of DNN models.

4.2 Convolutional Neural Network (CNN) Model

Convolutional neural network (CNN) is a kind of neural network model with local connection and weight sharing characteristics, which is suitable for processing local data such as images and speech. In voice arousal technology, CNN model can be used to extract local features from voice signals to improve the recognition accuracy of wake words.

(1) Principle of CNN model

The CNN model consists of convolution layer, pooling layer and fully connected layer. The convolution layer extracts the local features of the input data through convolution kernel, the pooling layer reduces the dimension and computation amount of the data through down sampling, and the fully connected layer classifies and recognizes the extracted features. The CNN model is trained by back propagation algorithm to constantly adjust the weights between convolutional nuclei and neurons to minimize classification errors [7].

(2) The application of CNN model in voice arousal

In voice wake technology, CNN model can be used to construct feature extractors and classifiers. Firstly, the pre-processed speech signal is converted into a two-dimensional representation such as a spectrum diagram or a Mayer spectrum diagram, and then input into the CNN model for feature extraction and classification. The CNN model extracts local features from speech signals through convolution layer and pooling layer, and then classifieds and recognizes them through full connection layer. By setting a threshold, you can determine whether the input voice signal contains a specific wake word.

(3) Optimization of CNN model

Aiming at the resource limitation problem in the smart home scene, the CNN model can be optimized. First, lightweight network structures, such as MobileNet and ShuffleNet, can be used to reduce the number of parameters and the amount of computation in the model. Secondly, techniques such as depth-separable convolution can be used to reduce the computational complexity of the convolution layer. In addition, techniques such as model pruning and quantization can be used to further compress and optimize the performance of the CNN model (see Table 1).

Table 1: CNN model optimization techniques

Optimization technique	Peculiarity
Lightweight network architecture	Reduce the number of parameters and calculation of the model
Depth-separable convolution	Reduce the computational complexity of the convolutional layer
Model pruning	Remove convolution kernels and joins that have less impact on model performance
Quantification	The weight and activation values of the model are converted from high precision to low precision, reducing the storage and computation requirements of the model
Model compression	A variety of methods (such as pruning, quantization, etc.) are used to compress the model to further reduce the complexity and power consumption of the model

5. Low Power Design Research

5.1 Event-driven Wake up Mechanism

Event-driven wake up mechanism is a low-power design method that triggers the wake up process by detecting key events in the speech signal. The method maintains a low power state when the voice signal does not arrive, and wakes the system for processing when the voice signal is detected. This approach can significantly reduce the power consumption of the system while maintaining the ability to respond to speech signals in real time.

(1) Event detection algorithm

Event detection algorithm is the core of event-driven wake mechanism. Commonly used event detection algorithms include energy detection algorithm, zero crossing rate detection algorithm and so on. The energy detection algorithm calculates the energy of speech signal to determine whether there is speech signal. The zero crossing rate detection algorithm determines whether there is a speech signal by calculating the frequency of the speech signal passing through the zero point [8]. By setting the threshold, you can determine whether the input voice signal meets the wake-up conditions (see Table 2).

Table 2: Event detection algorithm and characteristics

Algorithm name	Description	Peculiarity
Energy detection algorithm	The energy of speech signal is calculated to determine whether there is speech signal	Simple and easy, sensitive to energy changes of speech signals
Zero crossing rate detection algorithm	By calculating the frequency of the speech signal passing through the zero point to determine whether there is a speech signal	Sensitivity to the frequency characteristics of speech signals helps distinguish speech from noise

(2) The realization of wake-up mechanism

The event-driven wake up mechanism can be implemented by low power hardware accelerators. First, the event detection algorithm is embedded in the hardware accelerator to realize real-time event detection. When a voice signal is detected, the hardware accelerator triggers the wake-up process and transmits the voice signal to a subsequent processing module for processing. By optimizing the design and implementation of the hardware accelerator, the power consumption and response time of the system can be further reduced.

5.2 Low Power Hardware Accelerator

Low power hardware accelerator is a kind of hardware circuit specially used to accelerate a specific algorithm or task, which has the characteristics of low power consumption and high performance. In smart home scenarios, low-power hardware accelerators can be used to accelerate the implementation of voice wake-up algorithms to improve system performance and power efficiency [9].

(1) Design of hardware accelerator

The design of hardware accelerator needs to consider the characteristics of the algorithm and the limitation of hardware resources. First of all, it is necessary to decompose and optimize the voice wake-up algorithm, determine the key steps and modules that need to be accelerated, and then design appropriate hardware circuits and architectures according to the limitations of hardware resources and power consumption requirements [10]. In the design process, it is necessary to fully consider the parallelism and pipeline characteristics of hardware circuits to improve computing efficiency and reduce power consumption.

(2) Implementation of hardware accelerator

The realization of hardware accelerator can adopt digital signal processing (DSP), field programmable gate array (FPGA) and other technologies. DSP is a chip specially used for digital signal processing, which has the characteristics of high performance and low power consumption. An FPGA is a programmable logic device that can be configured into different circuits and functions as needed. Efficient hardware accelerator design and implementation can be realized by selecting suitable hardware platform and development tools.

6. Conclusion

In this paper, low resource voice wake-up technology in smart home scene is studied deeply, and a method based on low power feature extraction and neural network optimization is proposed. This method not only guarantees the wake rate, but also effectively reduces the power consumption and computational complexity, and is suitable for the practical application of smart home devices. The research of this paper provides a new idea and method for voice wake-up technology in smart home system, which is of great significance for promoting the development and application of smart home system. In future research, we should continue to explore new low-power feature extraction algorithms and neural network model optimization methods to further improve the performance and power efficiency of voice wake-up technology in smart home scenarios.

Acknowledgements

This work was financially supported by Guangdong Neusoft Institute 2023 University-level research project: Research on Low-Resource Voice Arousal Technology in Intelligent Home Appliances (2023DRXJ020).

References

[1] Baidu Online Network Technology (Beijing) Co. Ltd. Patent Application Titled "Method And Apparatus For Waking Up Via Speech" Published Online (USPTO 20200328903)[J]. Technology & Business Journal, 2020, 983.

[2] Wireless Communication Companies. Patent Issued for Voice Recognition Function Realizing Method and Device (USPTO 9542935)[J]. Journal of Engineering, 2017, 1858.

[3] Muda L, Begam M, Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques[J]. CoRR, 2010, 4083

[4] Qiuying S, Shiwen D, Jiqing H. Task-driven common subspace learning based semantic feature extraction for acoustic event recognition[J]. Expert Systems With Applications, 2023, 234.

[5] Hoy M B. Alexa, Siri, Cortana, and more: an introduction to voice assistants[J]. Medical reference services quarterly, 2018, 37(1): 81-88.

[6] Chen G, Parada C, Heigold G. Small-footprint keyword spotting using deep neural networks[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, 4087-4091

[7] Stafylakis T, Tzimiropoulos G. Zero-shot keyword spotting for visual speech recognition in-thewild[C]. Proceedings of the European Conference on Computer Vision, 2018, 513-529.

[8] Dudley H, Balashek S. Automatic recognition of phonetic patterns in speech[J]. The Journal of the Acoustical Society of America, 1958, 30(8): 721-732.

[9] Weintraub M. Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system[C]. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993, 2: 463-466

[10] Wang Y, Long Y. Keyword spotting based on CTC and RNN for Mandarin Chinese speech[C]. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, 374-378