Customized Wake-Up Word with Key Word Spotting using Convolutional Neural Network

Tsung-Han Tsai
Department of Electronic Engineering
National Central University
Chung-Li, Taiwan
han@ee.ncu.edu.tw

Ping-Cheng Hao
Department of Electronic Engineering
National Central University
Chung-Li, Taiwan
105521050@dsp.ee.ncu.edu.tw

Abstract— **In this paper, a customized wake-up word system combined with key word spotting using neural network was proposed. This system is divided into three phases: training wake-up word phase, detecting wake-up word phase and key word spotting phase. In training phase, user can say any word in any language and system will automatically count how many syllable of this word. If several syllables are in the range, system will accept this customized wake-up word. Next, the word will be extracted the features by Mel-Frequency Cepstral Coefficients (MFCC) method. It can be used for speaker model, speech model and state sequence for next phase. In detecting phase, system detects an unknown voice segment and compares it with models. After these steps, system will determine to wake up or not. If user says the right wake-up word, system goes to next phase. In key word spotting phase, the command words are fixed. The system is designed using convolutional neural network for key word spotting model. Moreover, all processes are executed without Internet to protect user privacy. This system can give a good result with a very small amount of wake-up word training data, and run in real-time.**

Keywords: customized wake-up-word; mel-frequency cepstral coefficients; gaussian mixture model; hidden markov model; convolutional neural network;

I. Introduction

Thanks to new technological advances, intelligent voice assistant becomes more and more popular. It helps us to do lots of things without hand, such as controlling the light or getting important information. Most of device need a wake-up word to activate the service. However, the word is fixed and cannot be changed. If someone wants to wake it up, he/she must say the words set by the company, which is not feeling free for consumers. For example, only speaking “Hey Siri” can wake Apple voice assistant up.

The proposed system let users set their wake up word in any language. It means everyone can set their own word. The system will know who is speaking now and then find out the word is right or wrong. If wake-up action is successful, user can say the specific control command to do something. The control commands are fixed. To customize the wake-up word, user just say the word and the system will automatically judge if this word is between 3 and 6 syllables or not. If not, the system will ask user to change another one. This is because 3 to 4 syllables of wake-up word is the best choice in the research. Wake-up word less than 3 syllables will increase False Alarm (FA) rate/False Accept rate, and more than 6 syllables will decrease accuracy/True Accept rate. All of the processes are executed without Internet, so that the user’s sound data will be privacy and safety.

We use Voice Activity Detection (VAD) to cut the voice, use Mel-Frequency Cepstral Coefficients (MFCC) to extract the feature, use Gaussian Mixture Model (GMM) to make the speaker identification model, use Gaussian distributed Hidden Markov Model (HMM) to make word model and use Convolutional Neural Network (CNN) to make command word model.

II. THE PROPOSED SYSTEM

The proposed system flowchart is provided in Fig 1. First, the system will detect the voice segment by VAD with the 8000 sample rate and cut every 20ns for one frame, where the overlap is 10ns. Every frame will be extracted features by 20-dimension MFCC and 20-dimension first order differential. Next, it will use this data to make HMM with 6 state component and calculate the state sequence. If the length of state sequence is between the thresholds, the system will save this model and sequence to the data pool. Also the system will use same feature data to train a 16-component GMM and combine it with the HMM.

After record the wake-up word, it will go into detection part. In this part, the system uses the same method to detect an unknown voice segment and extracted its features. Then it uses these features to calculate likelihood with each GMM, finding out the most likely speaker. Next, it will also use these features to get the state sequence through corresponding HMM and predict posterior probability of each GMM component. If both state sequences are mostly the same and the posterior probability pass the threshold, it means waking-up action is successful, and then user can say the command word to do something. Otherwise, the system will ignore this voice segment and prepare to detect next one.

The part of command word detection is also called Key Word Spotting (KWS). We use CNN to train some fixed words. The structure is described in [1]. In this work, “Yes”, “No”, “One” and “Two” were been chosen. There are four main modules in this system: MFCC, GMM, HMM and CNN. Each will be introduced below.

978-1-7281-2478-0/19/$31.00 ©2019 IEEE

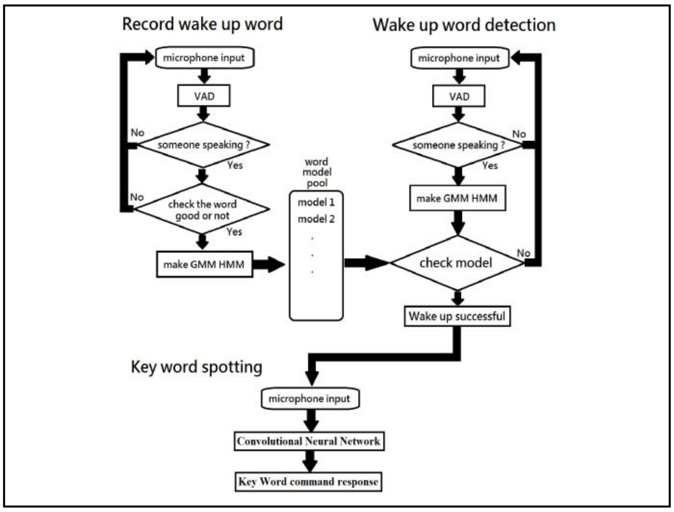

Fig.1 system flowchart

A. *Mel-Frequency Cepstral Coefficients (MFCC)*

MFCC was proposed by Davis and Mermelstein and is a commonly used for extracting feature in speech processing in [2]. In human auditory perception experiments, human hearing only focus on certain frequency domains, like a filter bank. However, the filter banks are not evenly distributed in the frequency domain. There are very dense filters in low frequency, while sparsely distributed in high frequency. MFCC simulates the human's ear so that we can analyze features easily.

B. *Gaussian Mixture Model (GMM)*

GMM is a statistical model of distribution, composed of multiple single Gaussian probability density functions. The GMM can smoothly approximate the distribution of any shapes. In [3], the author used GMM to train speaker model. It turned out that this method gave the highest accuracy comparing with others, letting GMM be one of the main methods for speaker verification

C. *Gaussian distributed Hidden Markov Model (HMM)*

HMM is a statistical model used to describe a Markov chain process with hidden unknown parameters. It is often used in time series related problem such as dynamic image recognition and speech recognition. In linguistics, we can divide human speech into various syllables, and HMM is the best choice to do this. The details of every step were presented in [4].

D. *Convolutional Neural Network (CNN)*

CNN is a feedforward neural network with each neuron corresponds to a surrounding range. It is composed of one or more convolutional layers and several full connected layers, also includes pooling layer. The advantages of convolutional neural networks are: fewer parameters, high precision, and many improved structures. In [1], a three-layers CNN was proposed for small-footprint keyword spotting and had a good result.

III. EXPERIMENTAL RESULTS

In Fig. 2, we implement the system on Jetson TX2 (A) and use a microphone input (B). We test the wake-up word "Hello smart light" and Table 1 shows the result.

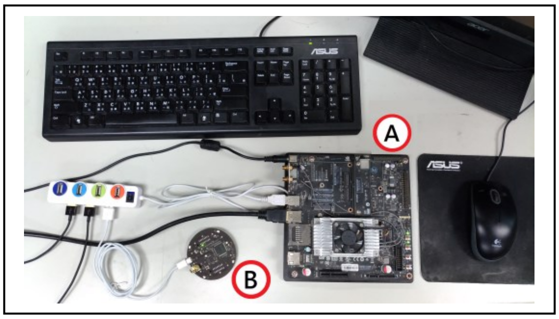

Fig.2 system implementation

TABLE I. ACCURACY, FALSE ALARM AND CALCULATION TIME OF WAKE-UP WORD

Accuracy	False Alarm	Calculation Time
99.6 %	3 times in 24 hours	0.2 second

IV. CONCLUSION

This paper proposed a customized wake-up word with key word spotting in real-time system. It can record any language customized wake-up word and count how many syllables are. The first step is using Gaussian Mixture Model likelihood to find out the true speaker. The second step is predicting posterior probability of each Gaussian Mixture Model component. The third step is compare the state sequence through Gaussian distributed Hidden Markov Model. If pass, the word model will be made. Next, system will detect sound and compare with models. If the word is correct, system will continue to detect command word. Otherwise the system will go to previous step to detect wake-up word. The simulation result shows the accuracy is 99.6%, False Alarm is 3 times per day and Calculation Time is 0.2 second.

REFERENCES

[1] Tara N. Sainath and Carolina Parada, " Convolutional neural networks for small-footprint keyword spotting," in INTERSPEECH, 2015.

[2] Davis and Mermelstein, " Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences" , IEEE transactions on acoustics, speech, and signal processing, VOL. ASSP-28, NO. 4, August 1980

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[4] Povey, Daniel et al, " The Kaldi Speech Recognition Toolkit ", IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US.