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

*Abstract***—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.**

**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.**

**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

The proposed system let users set their wake up word in any language.

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.

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.

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

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.

After record the wake-up word, it will go into detection part.

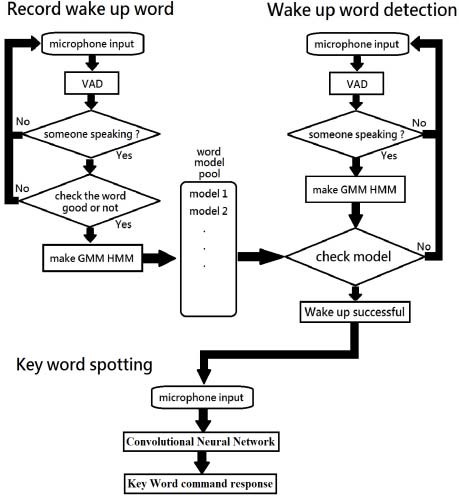
same method to detect an unknown voice segment and extracted its features. Then it uses these features to calculate likelihood with each GMM, finding out the most likely speaker.

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.

There are four main modules in this system: MFCC, GMM, HMM and CNN. Each will be introduced below.

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)

composed of multiple single Gaussian probability density functions. The GMM can smoothly approximate the distribution of any shapes

## 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.

## D. Convolutional Neural Network (CNN)

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.