ICIC Express Letters
Part B: Applications
Volume **7**, Number **9**, September **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1901–1909**

APPLICATION OF HMM-BASED CHINESE SPEECH RECOGNITION ON INTERNET OF THINGS FOR SMART HOME SYSTEMS

Neng-Sheng Pai[1,*], Shi-Xiang Chen[1], Pi-Yun Chen[1]
Chao-Lin Kuo[2] and Hong-Yu You[1]

[1]Department of Electrical Engineering
National Chin-Yi University of Technology
No. 57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung City 41170, Taiwan
*Corresponding author: pai@ncut.edu.tw

[2]Department of Maritime Information and Technology
National Kaohsiung Marine University
No. 482, Jhongjhou 3rd Rd., Cijin Dist., Kaohsiung City 80543, Taiwan

Received February 2016; accepted May 2016

Abstract. *We aim to develop a Chinese speech recognition system to apply to the control of Internet of Things (IoT) for smart home systems based on hidden Markov model (HMM). A speech reference model is established with the pre-processed speech signals through the characteristic parameters of Mel-frequency cepstral coefficients (MFCC), and this reference model will compare the user-input speech signals and achieve speech recognition. The speech recognition system makes its comparison with the use of hidden Markov model. Finally, this Chinese speech recognition system is applied to Internet of Things for smart home systems to allow the users to control them via speech recognition and achieve the effect of human-computer interaction to validate the reliability and viability of the system.*

Keywords: Chinese speech recognition, Internet of Things, Hidden Markov model, Mel-frequency cepstral coefficients

1. **Introduction.** The modern speech recognition systems can be traced back to the Automatic Speaker Isolated Digits Recognition System proposed by Davis et al. of the Bell Labs in 1952 [1] which extracted the spectral characteristics of vowels in a speech for recognition. In the 1980s, speech recognition development reached a peak, and the speech recognition method transformed from a rule-based model to a statistics based model. The hidden Markov model (HMM) [2,3] proposed by Baum was the most significant breakthrough in speech recognition. The HMM research made a substantial progress in the development of the continuous speech recognition system. At present, statistics-based models of the speech recognition system are the mainstream [4]. Common statistics-based models of the speech recognition system mainly consist of speech pre-processing, characteristic extraction, statistical acoustic models and language models. In the speech recognition system and the characteristic extraction, the main task is to extract from speech signals useful characteristic parameters. Most of the common characteristic extraction methods currently are used based on Fourier transform, linear prediction and cepstral analysis. Among them, linear prediction coefficients (LPC) [5] and Mel-frequency cepstral coefficients (MFCC) [6] are the ones that are most widely used. In the aspect of acoustic models, most systems use HMM as a basis. In 2012 Mohamed et al. proposed a pre-trained method in Deep Neuron Networks (DNN) [7] to solve the local convergence and computational intensity problems. However, the fact that deep learning requires substantial speech data as training samples makes it difficult to be used in the laboratory. Therefore, we use HMM-based speech recognition method. Our purpose is

to develop a controller based on Chinese speech recognition [8], as a bridge of communication between persons and Internet of Things. In this way, we can significantly reduce the complexity of the operation. A variety of electrical devices can be directly controlled by speech commands. It can be like interpersonal communication between people and devices and it can enhance the experience of human-machine interface.

2. Speech Characteristic Extraction.

The characteristics of a speech signal must be obtained before speech recognition. This chapter discusses how to obtain a speech signal's characteristic parameter.

2.1. Speech signal pre-processing.

Speech signals need to be pre-processed to extract the desired feature vector. Speech signal pre-processing contains audio sampling, pre-emphasis, frame blocking and windowing.

2.2. Mel-frequency cepstral coefficients.

The feature vector of speech signals must be extracted before speech recognition, and the feature vector is used to represent the characteristics of the speech signals. As signal characteristics cannot be easily identified from signal changes in the time domain, we use Mel-frequency cepstral coefficients (MFCC) to extract the characteristics.

2.3. Triangular bandpass filter.

The paper uses the Mel Scale to represent the human ear's sensory judgment of pitch changes, and the conversion equation between frequencies and the Mel Scale can be expressed in Equation (1).

$$mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Figure 1 shows the frequency at low frequencies, the spacing of triangular filters is dense and bandwidths are narrow. And with the increase of frequencies, the spacing and bandwidths also increase in order to simulate the human ear which has better sensitivity at low frequencies than high frequencies.

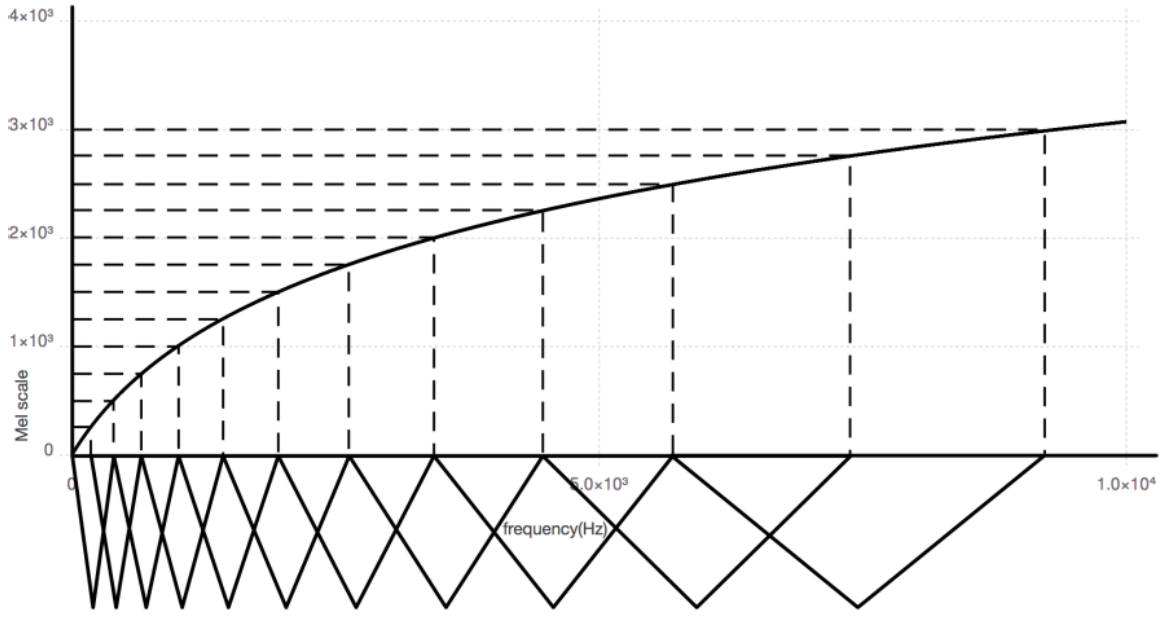

FIGURE 1. The corresponding relationship between frequencies and the Mel-frequency filter group

3. **Speech Recognition Based on HMM.** Speech recognition can be divided into isolated-word speech recognition and continuous speech recognition according to different targets of recognition. Because isolated-word speech recognition uses dynamic time warping (DTW), each speech input needs to be compared with all of the reference speech samples. If the vocabulary volume is large, the time required for recognition is long. Besides, because a specific speaker's voice is used as reference samples, the system's speech recognition effect is poor other than the specific person's speech, and thus a universal speech recognition system cannot be established. In order to accelerate the speed of speech recognition and establish a speaker-independent speech recognition system, we use continuous speech recognition based on the hidden Markov model to avoid these problems.

3.1. **Acoustic model.** In HMM-based speech recognition, an acoustic model needs to be established first. Because continuous speech recognition is required, a phoneme is used as a speech recognition unit, and an HMM is used to describe each unit to be recognized. Acoustic model training is the estimation of HMM's μ and Σ. Every state output of HMM can be expressed as an independent Gaussian probability density function as in Equation (2). $b_j(o_t)$ is the state probability, showing the probability of b_j in the o_t state.

$$b_j(o_t) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} e^{-\frac{1}{2}(o_t-\mu_j)'\Sigma_j^{-1}(o_t-\mu_j)} \tag{2}$$

The maximum likelihood estimation of μ_j and Σ_j can be expressed as in Equation (3):

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)}, \quad \hat{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t-\mu_t)(o_t-\mu_t)'}{\sum_{t=1}^{T} L_j(t)} \tag{3}$$

where $L_j(t)$ represents the probability of being in state j at time t. We use the Baum-Welch algorithm to find the parameters of the acoustic model.

3.2. **Baum-Welch algorithm.** Baum-Welch algorithm [9] is an algorithm for HMM model's parameter revaluation. In order to calculate $L_j(t)$, the forward probability $\alpha_j(t)$ of model M is defined in Equation (4).

$$\alpha_j(t) = P(o_1, \ldots, o_t, x(t) = j|M) \tag{4}$$

$x(t) = j$ means it is in state j at time t, and $\alpha_j(t)$ is the joint probability of the sum of t speech vectors in state j at time t. Equation (4) is represented as a recursive form as Equation (5) to make calculation efficient. Equation (5) can be illustrated in Figure 2.

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1) a_{ij}\right] b_j(o_{t+1}), \quad 1 < j \leq N \tag{5}$$

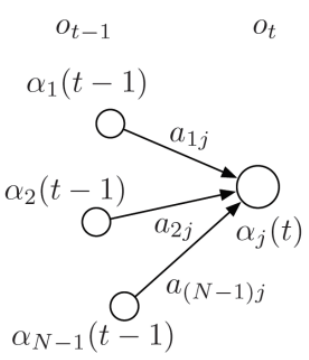

FIGURE 2. Forward probability graph

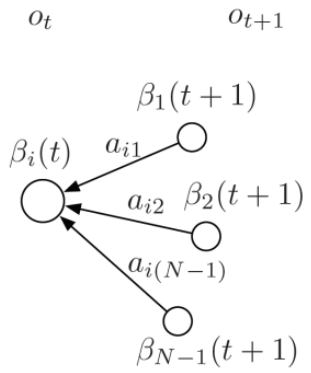

FIGURE 3. Backward probability graph

The termination condition Equation (4) similar to the forward probability defines the backward probability $\beta_j(t)$ of model M as in Equation (6).

$$\beta_j(t) = P(o_{t+1}, \ldots, o_T, x(t) = j|M) \tag{6}$$

The recursive conversion is also performed for backward probability, such as in Equation (7). Equation (7) can be illustrated in Figure 3.

$$\beta_j(t) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_i(t+1), \quad 1 \le j \le N \tag{7}$$

Therefore, the state of occupancy probability $L_j(t)$ can be obtained by calculating the product of the forward probability $\alpha_j(t)$ and the backward probability $\beta_j(t)$ as in Equation (8).

$$L_j(t) = \frac{1}{P} \alpha_j(t) \beta_j(t), \quad P = P(O|M) \tag{8}$$

The procedures of the Baum-Welch algorithm can be summarized as follows.

1) Calculate all the forward probability $\alpha_j(t)$ and backward probability $\beta_j(t)$ in state j at time t.

2) For each state j and time t, we use state of occupancy probability $L_j(t)$ and the current observation vector to update the density function μ and the covariance matrix Σ in that state.

3) Use the final density function μ and covariance matrix Σ to recalculate the model's parameter value.

4) If the $P(O|M)$ result of this iteration is not higher than the previous value, then stop the iteration; otherwise repeat the above calculation with a new parameter value.

3.3. **Recognition.** We use the Viterbi algorithm to search for the model which generates the maximum $P(O|M)$.

3.4. **Viterbi algorithm.** For a trained model, make $\phi_j(t)$ represent the maximum probability of the observed speech vectors $o_1, \ldots, o_t$ in state j at time t, and this can be calculated with the recursion equation in Equation (9).

$$\phi_j(t+1) = \max_{1 \le i \le N} \{\phi_j(t) a_{ij}\} b_j o_{t+1}, \quad 1 \le j \le N \tag{9}$$

The maximum likelihood $\hat{P}(O|M)$ is:

$$\phi_j(t+1) = \max_{1 \le i \le N} \{\phi_j(t) a_{iN}\}, \quad 1 \le j \le N \tag{10}$$

As in Figure 4, Viterbi algorithm [10] is the best method to find a path in a matrix. In the path from left to right at time t, each part of the path $\phi_j(T)$ for all the state j is known, so Equation (10) can be used to calculate the probability of the entire path.

3.5. **Continuous speech recognition.** The phoneme recognition is extended to continuous speech recognition. The trained acoustic models need to be connected in sequence, and each model in the sequence will correspond to a phoneme, as shown in Figure 5.

3.6. **Tri-phone model.** In order to increase the speech recognition accuracy, we add a tri-phone model in the related context. A tri-phone model uses a continuous tri-phone as the acoustic model, and the training of the tri-phone model is shown in Figure 6. A trained single-phoneme HMM is used to establish an initial tri-phone model through artificial markings of the table of tri-phone and phone. Furthermore, the initial tri-phone model and speech characteristics as well as artificial markings are retrained with the previously mentioned Baum Welch algorithm, and the state distributions among the models are connected with tree-based classifiers to make the output distribution more robust. After the process we can get the tri-phone HMM's model connects the states.

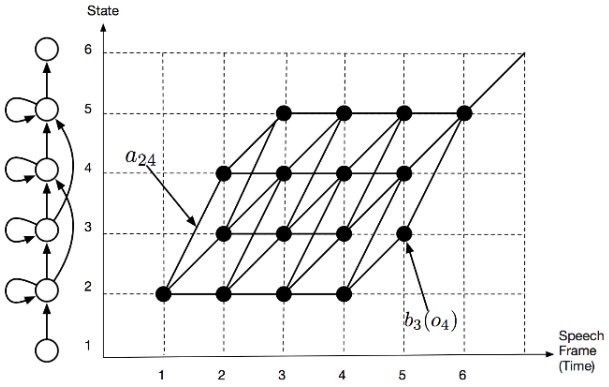

FIGURE 4. The signal recorder of the defect recognition system

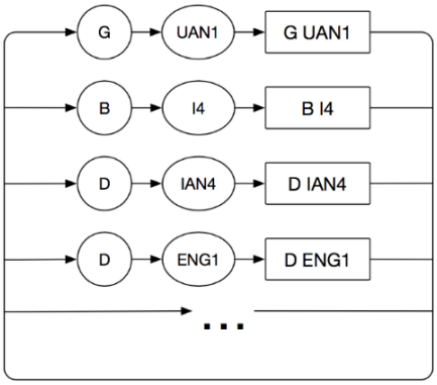

FIGURE 5. Continuous speech recognition

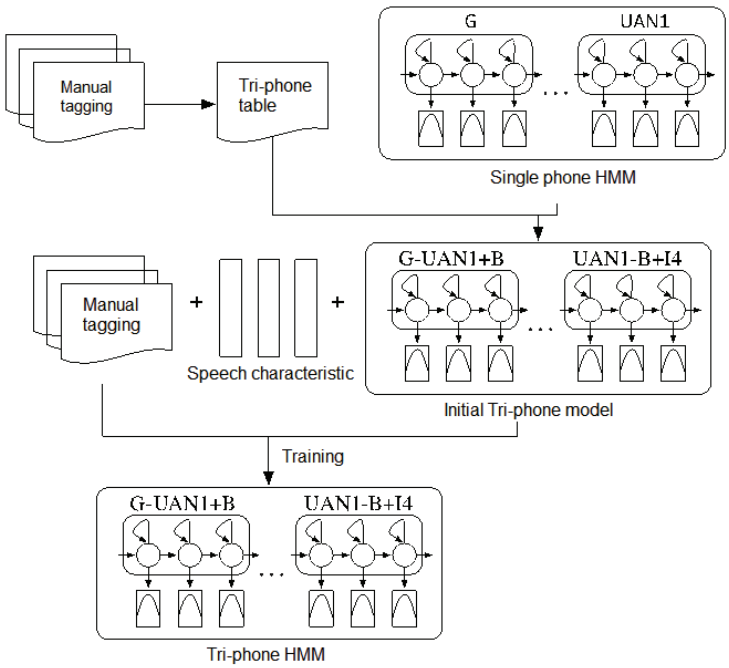

FIGURE 6. Tri-phone training

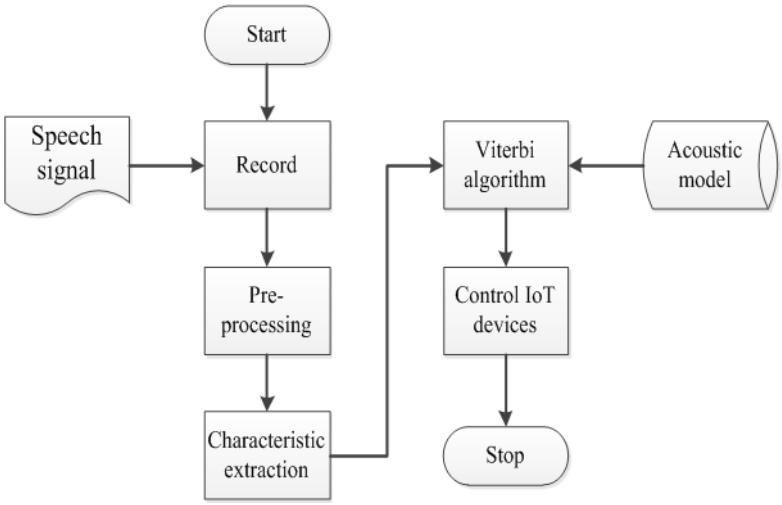

FIGURE 7. Speech recognition-based IoT for smart home system

4. **Experiment Results.** We design and implement a speech-controlled IoT for smart home system. The flowchart of system is shown in Figure 7 and the system architecture is shown in Figure 8. The client side mainly includes computers, smart phones and wearable equipment with a browser and microphone. HTML5, JavaScript and other Internet technologies are used to record a speech, and the speech is transmitted to the server side for analysis via the Internet WebSocket protocol. The server side according to the function can be divided into speech recognition and IoT control. After the speech signals are received from the client side via WebSocket, they are converted by an embedded system to a machine understandable command to control other IoT devices. Arduinos as well as traditional electrical devices can be used for the controlled devices via the IoT interface, and here the Arduinos used are ArduinoYún and ArduinoUNO. First of all the speech signal recorded from the client side is transmitted to the server through the Internet. They are converted by an embedded system to a machine understandable command to control the IoT terminal equipment through the Internet.

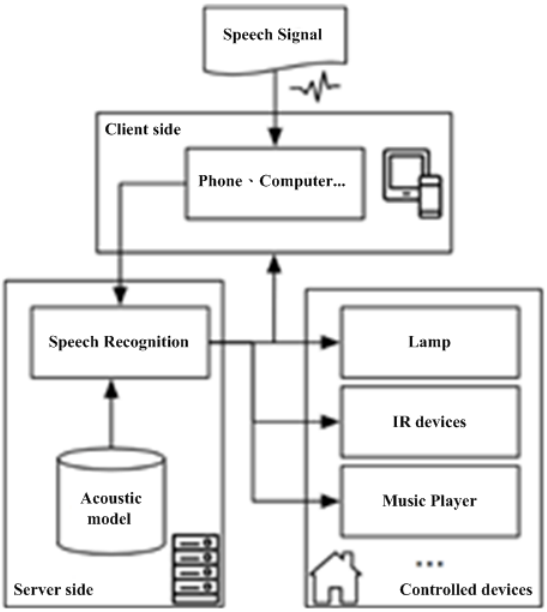

FIGURE 8. System architecture

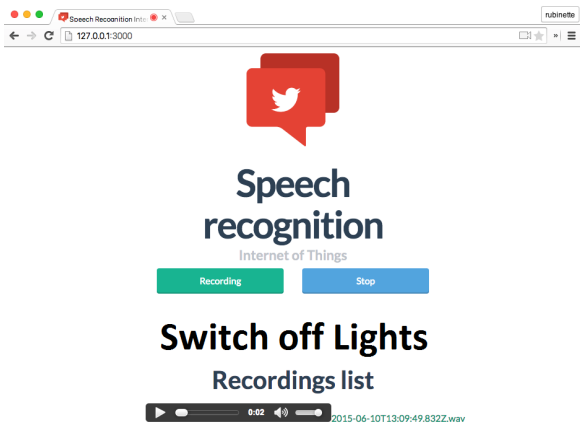

FIGURE 9. Recognition results

4.1. **Speech recognition test.** It requires a large amount of speech samples to train a model. We use 400 speech samples as the training set, and the recording specifications are mono, sampling frequency of 8kHz, sampling resolution of 8bits WAV files, and a quiet laboratory recording environment. After marking the speech content, the Baum-Welch algorithm mentioned in the paper is used for phoneme training. We use a webpage as the interface for user interaction. Figure 9 shows the recognition results after the user on the client side uses his/her own microphone to give control commands.

4.2. **Recognition rate.** From here, percent correct and percent accuracy are defined first. The definition of percent correct is in Equation (11).

$$\text{Percent Correct} = \frac{N - D - S}{N} \times 100\% \tag{11}$$

The definition of percent accuracy is in Equation (12).

$$\text{Percent Accuracy} = \frac{N - D - S - I}{N} \times 100\% \tag{12}$$

where N is the total number of test speeches. D (Deletion Error): the phoneme is in the correct results but not recognized. S (Substitution Error): the phoneme is incorrectly

recognized. *I* (Insertion Error): the phoneme is not in the correct results but that is recognized.

4.3. **Test 1.** In Test 1 a recognition rate test is conducted on the number of frame sampling points and whether tri-phone recognition is used. 100 specific samples were used for testing, and the best parameters are selected for the system. Table 1 shows the correct recognition rates for phonemes and words of the respective frame.

Table 1. Frame sampling point numbers and overlap numbers

Frame size	Frame displacement	SENT	WORD
20ms	10ms	Correct: 98.00%	Correct: 97.26% Accuracy: 97.26%
25ms	10ms	Correct: 98.00%	Correct: 97.39% Accuracy: 97.39%
25ms	15ms	Correct: 98.00%	Correct: 97.26% Accuracy: 97.26%
30ms	15ms	Correct: 98.00%	Correct: 97.26% Accuracy: 97.26%

Table 2 shows the comparison of tri-phone-bound and non-tri-phone-bound recognition accuracy rates. Here we select the one with the best recognition effect, the frame with a length of 25ms and a displacement of 10ms as the main parameter, and use tri-phone to further improve the system's recognition rate.

Table 2. Whether Tri-phone is used

	Frame size	Frame displacement	SENT	WORD
Single phone	25ms	10ms	Correct: 98.00%	Correct: 97.39% Accuracy: 100.00%
Tri-phone	25ms	10ms	Correct: 100.00%	Correct: 97.39% Accuracy: 100.00%

4.4. **Test 2.** The comparison between specific speakers and non-specific speakers is shown in Table 3. The results show that the recognition accuracy of non-specific speakers is slightly lower than that of specific speakers, but the recognition accuracy is still good.

Table 3. The recognition accuracy between specific speakers and non-specific speakers

	Frame size	Frame displacement	SENT	WORD
Specific speakers	25ms	10ms	Correct: 100.00%	Correct: 100.00% Accuracy: 100.00%
Non-specific speakers	25ms	10ms	Correct: 90.00%	Correct: 98.63% Accuracy: 98.63%

4.5. **Test of IoT for smart home systems.** Figure 10 is the server interface, and Figure 11 is the webpage for a mobile phone. The user must use a computer, mobile phone or smart watch to transmit through the webpage the speech signals to the server, which will recognize and convert the speech signals into machine-understandable commands to control IoT devices. Figure 12 and Figure 13 show the results of a user giving a command to an IoT for smart home system through the webpage to switch on and off a light bulb.

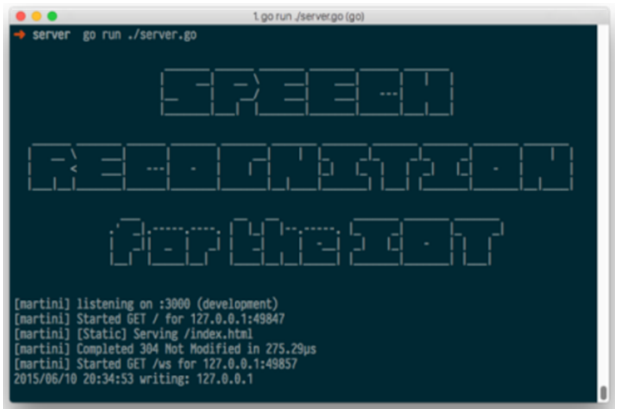

FIGURE 10. Server interface

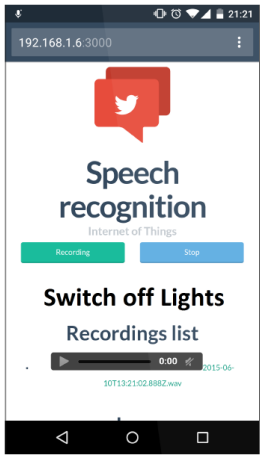

FIGURE 11. Mobile phone interface

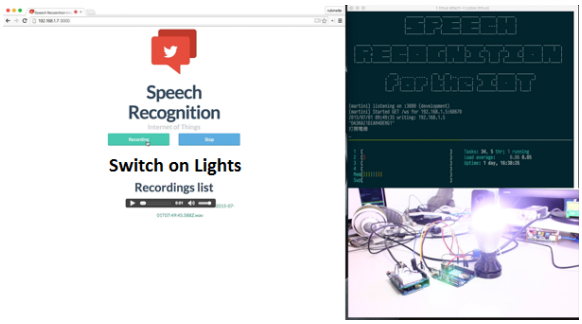

FIGURE 12. The light bulb being switched on

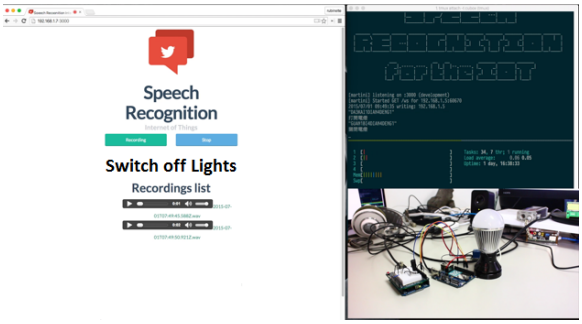

FIGURE 13. Recognition results

5. **Conclusion.** We complete an application of HMM-based speech recognition on Internet of Things for smart home systems. The system can be divided into two parts, namely speech recognition and IoT for smart home systems. Speech recognition mainly uses HMM and a tri-phone mode to establish a tri-phone-based acoustic model for speech signals, and uses the Viterbi algorithm to obtain a model closest to the input speech to achieve speech recognition. For non-specific speakers the recognition rate is as high as 90%. In order to reduce the cost of equipment replacement and let families enjoy the convenience brought by IoT for smart home systems. We use WiFi to transmit control command, and use embedded devices to provide the ability to connect IoT in an upgraded manner for the controlled device side. The future works also learn a lot of data to train the acoustic model by deep learning method, and furthermore improve the recognition rate. In terms of security, we can add the speaker recognition. Let IoT for smart home system with the ability identify a user.

REFERENCES

[1] K. H. Davis, R. Biddulph and S. Balashek, Automatic recognition of spoken digits, *Journal of the Acoustical Society of America*, vol.24, no.6, pp.637-642, 1952.
[2] S. Z. Yu, Hidden semi-Markov models, *Artificial Intelligence*, vol.174, no.2, pp.215-243, 2010.
[3] R. Su, X. Liu and L. Wang, Automatic complexity control of generalized variable parameter HMMs for noise robust speech recognition, *IEEE Trans. Audio, Speech, and Language Processing*, vol.23, no.1, pp.102-114, 2015.
[4] M. Padmanabhan and M. Picheny, Large-vocabulary speech recognition algorithms, *Computer*, vol.35, 2002.
[5] P. Vaidyanathan, *The Theory of Linear Prediction*, Synthesis Lectures on Engineering Series, Morgan & Claypool, 2008.

[6] V. Tiwari, MFCC and its applications in speaker recognition, *International Journal on Emerging Technologies*, vol.1, pp.19-22, 2010.

[7] A. Mohamed, G. E. Dahl and G. Hinton, Acoustic modeling using deep belief networks, *IEEE Trans. Audio, Speech, and Language Processing*, vol.20, no.1, pp.14-22, 2012.

[8] G. J. Jang, C. Pan, J. H. Park, J. S. Park and J. H. Kim, Recognition unit determination of interactive Chinese speech recognition for embedded devices, *IEEE Trans. Consumer Electronics*, vol.58, no.4, pp.1453-1458, 2012.

[9] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. of the IEEE*, vol.77, no.2, pp.257-286, 1989.

[10] A. Viterbi, A personal history of the Viterbi algorithm, *Signal Processing Magazine*, vol.23, no.4, pp.120-142, 2006.