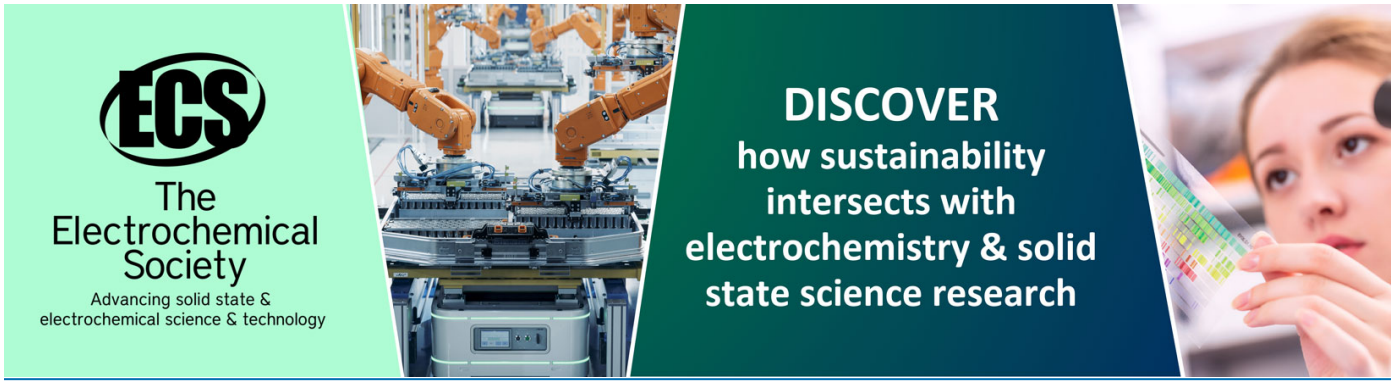
Application of Automatic Speech Recognition (ASR) Algorithm in Smart Home



IOP Conf. Series: Journal of Physics: Conf. Series **1237** (2019) 022133 doi:10.1088/1742-6596/1237/2/022133

Application of Automatic Speech Recognition (ASR) Algorithm in Smart Home

Hong Chen [1], Bo Zhang [2,*]

School of Computer Science, Wuhan Donghu University, Wuhan 430212, China

*Corresponding author's e-mail: bob.cheung@ovspark.com

Abstract. Study on practical ASR system for smart home had significant meanings for the development of the smart home. Through analysis of embedded ASR technology and the control technology of smart home, people could utilize VS1003 to record the speech for audio decoder chip by taking the NL6621 template as the platform. And, the ASR system on smart home could be realized by utilizing the Hidden Markov Model (HMM) algorithm to conduct speech model training and speech matching. Test showed that, the ASR system has relatively high recognition rate and real-time property.

1. Introduction

Smart home was mainly utilized to improve users' home environment and life quality. And, the ASR was the key to get rid of complicated remote manual control operation.

Pre-processing element was mainly used to sampling, quantify and code the speech signal, to realize the speech enhancement by utilizing the denoising algorithm of small wave; the function of pre-emphasis was to enhance the resolution ratio of high frequency; as the speech signal had short time steady property, so it could conduct segment treatment for speech signal, namely windowing and framing; you needed to have endpoint detection for input signal before feature extraction, which was the one important procedure of the key word "Sampling" in this paper. The purpose of almost all the endpoint detections for ASR system was to only detect the start time and end time of the speakers' speech. However, this paper proposed the detection method according to the endpoint classified by syllable, as was shown in figure 1:

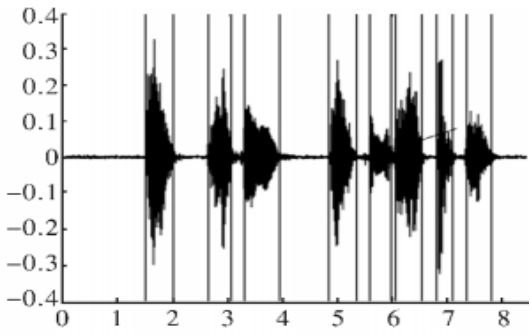


Figure 1 Detection results according to syllable

Regarding on the extraction element of the characteristic parameters, Mel Frequency Cepstrum Coefficient (MFCC) was utilized to extract these parameters, which could represent the basic characteristics of the speech; You could establish reference template library according to the

characteristic parameters; Regarding on pattern matching element, you could conduct characteristic extraction and similarity calculation with the characteristic parameters in the template library one by one after detecting the endpoint input by the users according to the syllable, in case of the similarity matching $P < Pmin$, directly ignored these as junk information, in case of $P > n$, you could identify it as the syllable of the key words or key words. The algorithm divided the key words in the smart home into multiple syllables (words), established reference template for each syllable during template training phase, and also established template for all the key words of the speech.

During recognition phase, you could match these templates classified by the endpoint of the syllables, when the first syllable of the key word was matched, then moved to the second one. If the continuously matched several words were the syllables in a certain key word, you could divide these continuous syllables and match with the template of the key words, if it was successfully matched, then it was confirmed such continuous syllables were the key word. The syllables were the key word.

2. MAIN ALGORITHM ON ASR

Currently, the commonly representative ASR algorithm included Dynamic TimeWarping (DTW), Vector Quantization (VQ), Hidden Markov Model (HMM), Artificial Neural Networks (ANN), etc.

A. DTW

In order to solve the problem having various durations when directly matching speech parameters with template parameters, Japanese scholar, Itakura, proposed a dynamic timewarping technology.

The idea of this algorithm was to non-uniformly warp or bend the timer shaft of these speech signals ready to be identified, in order to make its characteristics be aligned with the template characteristics, and continuously calculate the minimum matching path of the distance between the two vectors, so as to get the minimum dynamic timewarpping function.

As the speech speed was not uniform, utilization of DTW could solve such problem, which was applied widely in early ASR system.

DTW was a nonlinear dynamic timewarpping technology, which mainly combined the timewarpping with distance measurement.

Assume that:

(1)A={al，a2，…am,...,aM} was the sequence of the characteristic vectors, with totally M frames in amount, am was the speech characteristic vector of the number m;

(2) B={bl，b2，…bn,...,bN}，(M ≠N) was the input sequence of the characteristic vector, with totally N framing vector input, bn was the speech characteristic vector of the number n.

$$D = \min_{w(n)} \sum_{n=1}^{N} d[n, w(n)]$$

Figure 2 showed the DP algorithm. In a 2 dimensional rectangular coordinate system, horizontal axis represented each frame number of the input template as n=1-N, and longitudinal axis represented each frame number of the reference template as m=1-M.

The integer coordinate lines of the horizontal and longitudinal axis formed a grid, and each crossing point in this grid represented the intersection between one frame in testing template and one frame in reference template.

DTW algorithm included 2 parts: 1. Calculate the distance of the frames in input and reference templates, and get the matched distance matrix of the frames; 2. Obtain a optimal path in the matched distance matrix of the frames.

The accumulated distance of the path:

D(n, m)=d(n,m)+min[D(n-1, m-1), D(n-1, m-1), D(n-1, m-2)]

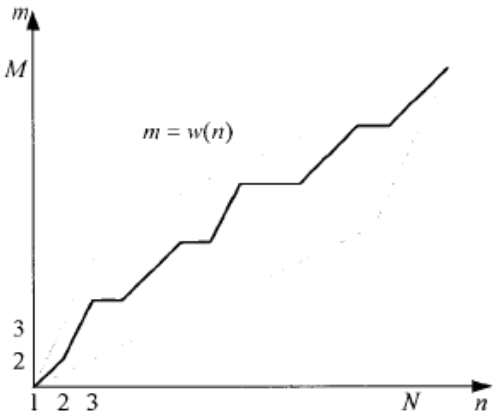


Figure 2 Block diagram on algorithm function

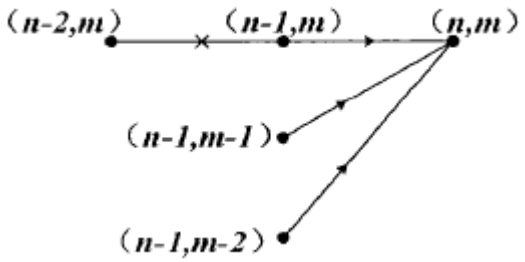


Figure 3 Continuous conditions

B. Hidden Markov Model (HMM) template

HMM algorithm was a numerous speech data based statistical template, and currently it obtained wide application in speech signal treatment field. The basic theory and variously practical algorithms of HMM were the important foundations among the modern speech recognition. Large numbers of experiments showed, HMM could indeed describe the generating process of the signal precisely. As HMM algorithm involved in the training process, the obtained statistical template was relatively steady, enabling to handle various emergency situations. Therefore, HMM algorithm had good recognition rate and anti-noise performance.

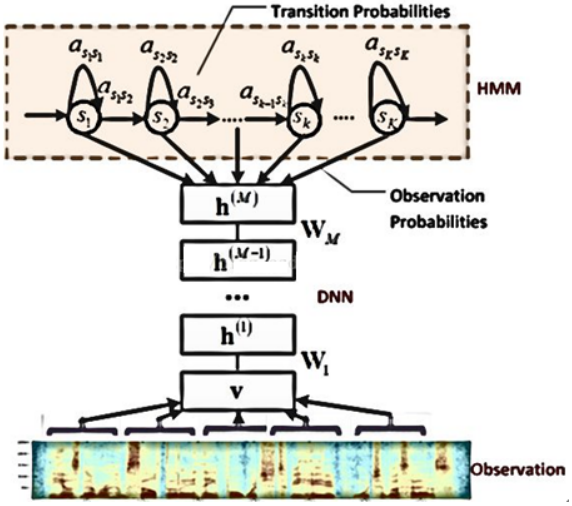


Figure 4 Relation between HMM and speech parameters

(1) Forward probability

The forward probability of HMM represented, the probability at state i in partial observation order {O1 , O2 , O3... Ot} when the time was t for the given HMM template.

Termination of calculation was shown in figure 5

From figure 5, the final state results were calculated by utilization of forward algorithm and the iteration of each state.

From figure 6, for forward algorithm on iterative calculation of each state from si to SN, you could observe the output probability of the order. The formulas on forward and backward probability divided the HMM template's output probability of the whole observation order into the output probability results of the two parts, and they all had their corresponding recursion formulas respectively, which could largely simplify the calculation. After analysis, the following formula on output probability was obtained:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \partial t(i)\beta t(i) = \sum_{i=1}^{N} \partial r(i), 1 \le t \le T-1$$

The topological structure of HMM could be represented visually by using Bayes network. As was shown in figure 7, it indeed expressed the HMM topological structure state visually. And, the detailed procedures on forward and backward iterative algorithms were given. We could learn from the definition of HMM, the occurred events were rightly the observed variable values, which were represented by the transparent circle symbol. It represented the observed value at the time Ot(1T), t but the implied state could not be observed directly.

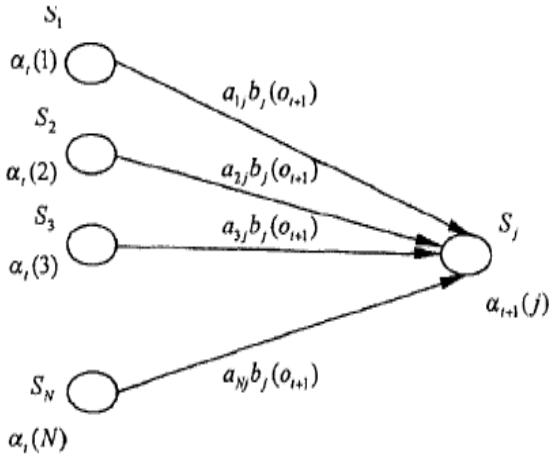


Figure 5 Schematic diagram of forward algorithm

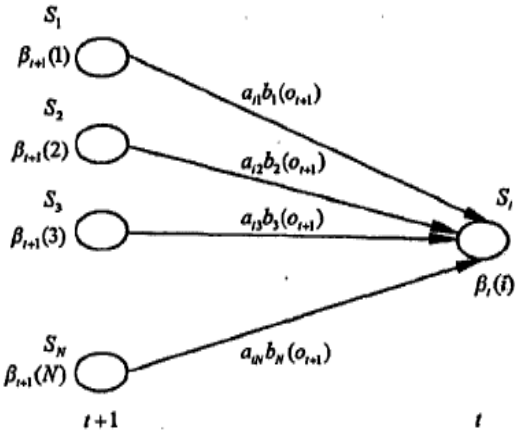


Figure 6 Schematic diagram of backward algorithm

The shadow circle symbol represented, the state at the time of t in case of qr = (1 ⩽T). The solid lines represented the relation between the events and the states. The dotted arrows represented HMM parameters could not be expressed directly, and they must experience the training process, so as to determine its real meanings.

The shortcoming on the training method of maximum mutual information was that it had not good iterative algorithms, which could not ensure convergence and not to cross the border on the repeatedly evaluated value, so the training method on MME was considered to be improved. Here, it was considered to conduct MME training by combining with genetic algorithm.

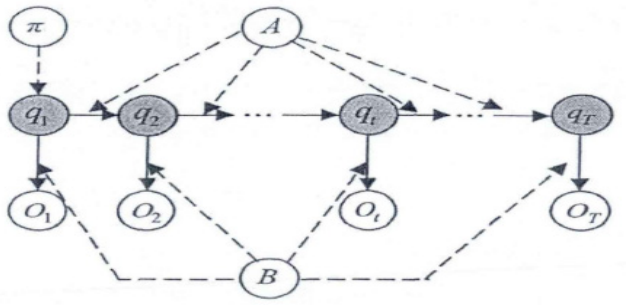


Figure 7 Baycs network diagram

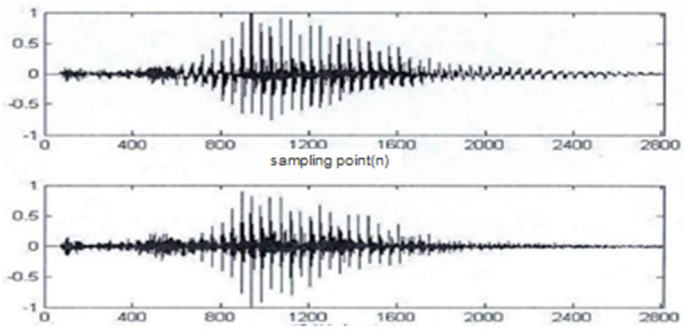


Figure 8 Waveforms of speech signals before and after pre-emphasis

From figure 8, the sharp noises could be well eliminated by comparing the waveforms of speech signals before and after pre-emphasis.

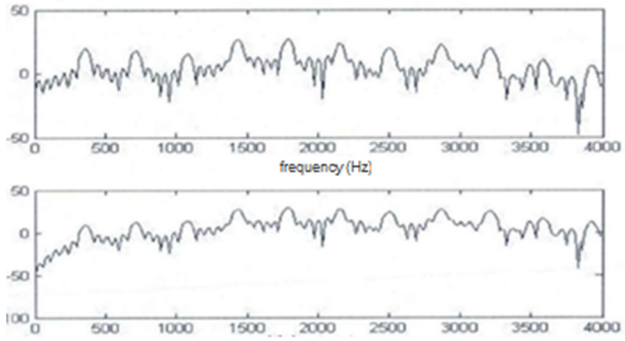


Figure 9 Frequency domain amplitude responded waveforms before and after pre-emphasis of frame data

Figure 8 was the comparison design sketch on various frequency domain amplitudes of frame data before and after audio order. From figure 9, the frequency domain amplitude of the frame data before and after pre-emphasis had a certain change, use of the improvement measure on MMIE training method and use of genetic algorithm combined method had very good effect, and the training template could relatively be reduced, as well as, training effect got significant improved.

Markov Model could also be treated as random finite-state. Markov chain could be represented as the uncertain (not sure) finite-state momentum sketch on the transferred arc, as was shown in figure 10:

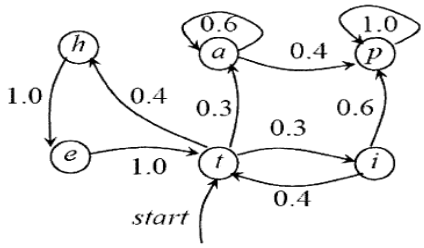


Figure 10 State sketch on Markov chain

3. APPLICATION OF ASR ALGORITHM IN SMART HOME

A. NL6621 based embedded hardware design

The hardware platform of speech recognition mainly included central processing unit NL6621, which could read and write memory, audio card chip vs1003 and some auxiliary equipment. Structure of hardware system was shown in figure 2. The main system utilized the NL6621, which was from Nufront corporation. The highest dominant frequency used by MCU was 160MHz, which could support 802.11b/g/n/i/e/p, Wi-Fidirect, BSSSTA, soft AP, WiFi safeguard setting and the security protocol of WMM-PS and WPA/WPA2. The chip of the codec was vs1003, with its data communication with NL6621 being realized through SPI trunk. It integrated the input interface of the microphone, the output interface of audio, and IMAADPCM coding conducted by voice input or line

input, which could efficiently accept and play audio information. Realization of hardware circuit: VS1003 was through the high or low pin value of the xCS and xDCS to confirm if the interface was at transmission state. The control command and data of NL6621 was received by SCI and SDI, the speech stream was obtained through SCI_HDAT1; VS1003 function control.

B. NL6621 based embedded software design

Software design mainly included 2 parts to realize the software-control embedded system and the compiling of HMM based ASR algorithm. The basic software architecture was shown in figure 11. Specific to the control part of the embedded system, it included the hardware initialization and collection of audio signals. It mainly utilized the software development kit provided NL6621, and utilized SDK to compile application program, including initialization of hardware pin, baud rate matching, configuration of the sound recording file, WiFi configuration, record, as well as, format conversion of audio file. Writing tool was needed after the completion of the program. After starting system, initialization of hardware module was needed firstly. Then, when the system started to operate, it was needed to collect speech by speech input equipment MIC and input speech by audio card VS1003. When the system detected speech input, speech recognition started to run, judging if the recognition is right or not.

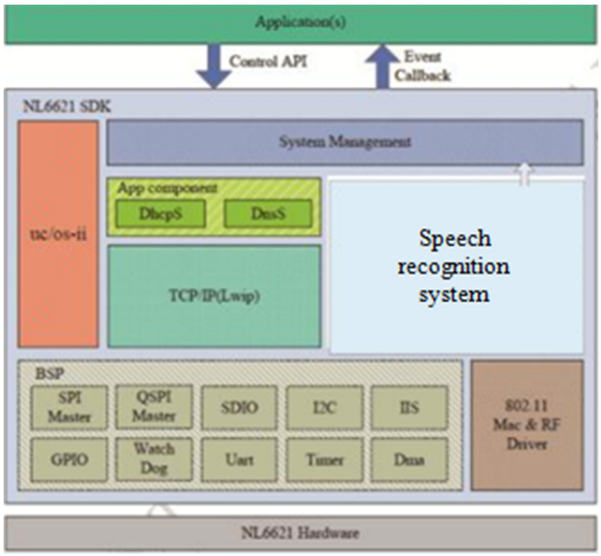


Figure 11 Structure chart on system software

4. Conclusion

Application of smart speech technology was an important part of AI industry, and construction on the open platform of smart speech technology based interface was an important realization approach, making AI hardware product and software services in a large-scale and systematized development. So, the enterprise could utilize the cloud platform to realize ASR and other AI functions, further popularize the smart phone, loudspeaker box, remote-control unit and other products of the customers in a rapid and large-scale way, which all in all quickly advanced the AI permeability rate and drove the development of AI industry.

Acknowledgments

This paper is supported by Scientific Research Project of Education Department of Hubei Province （B2018295） and the Foundation for Young Scholars Wuhan Donghu University under grant 2018dhzk006.

References

[1] Cao Shen, Chang Le. Home Service Design on A Smart Home Service Machine[J]. Application of singlechip and embedded system,2016,16(10):62-66.

[2] Yang Yefen, Ye Chengjing. A GSM Based Smart Home Speech Control System [J]. Computer System Application,2017,26(02):68-72.
[3] Zhang Pengyuan, Ji Zhe, Hou Wei, Jin Xin, Han Weisheng. Design and Optimization on ASR Algorithm under Small Resource Conditons [J]. Tsinghua University Report (natural science edition),2017,57(02):147-152.
[4] Jiang Tai, Zhang Linjun. Application of Adaptive Algorithm of ASR in Smart Home [J]. Computer System Application,2017,26(03):150-155.
[5] Zhang Shuailin. An Improved Key Word Recognition Algorithm on Smart Home Speech [J]. Electronics Technology,2017,30(07):5-8.
[6] Jing Niqin. Research on Smart Speech Module Based Smart Home System [J]. Electronic Production,2018(Z1):25-27+14.