ATLANTIS PRESS

Advances in Computer Science Research (ACSR), volume 81

2017 International Conference on Computational Science and Engineering (ICCSE 2017)

Design and Development of Speech Recognition System Based on HMM Algorithm

Xu Zhengru

Qingdao Huanghai University

Keywords: Markoff model, HMM, Speech recognition

Abstract: This thesis introduces the basic theory of speech signal processing, and reviews the development history of speech recognition both at home and abroad. The thesis also expounds the basic theory of Dynamic Time Warping Algorithm and Markov Model in detail, and studies the method of their application to small vocabulary recognition for language for specific people. After being checked, there is a experiment on speech recognition algorithm for Chinese small vocabulary for specific people and the results show that the improved speech recognition method has better recognition performance than the traditional one.

Introduction

Speech recognition is a high technology for a machine to translate a speech signal into a corresponding text file or command through identification and understanding. This thesis presents the occurrence, development and evolution progress of speech recognition research, research emphases in different periods, great achievements and even academic thoughts. The principle of speech signal recognition is to build sound tube model and its mathematical expression by studying the physical process of human voice[1]. The theory and practice development of speech signal management built from the physical mechanisms and processes of pronunciation has been very mature[2], but the listening system like the human's hasn't been developed. Another type of study that is being carried out is the acoustic theory that focuses on the perception of auditory physiological organs and auditory psychology on the speech, but little progress has been made. Mature theories can encode, compress, transfer, recognize, synthesize, and broadcast speech.

Originally, the thesis is based on searching for a large number of speech recognition technology products and through the research and analysis of the collected documents, here comes the conclusion that because the speech signal is time-varying and non-stationary, it is not ideal to apply the product to life due to the influence of the recognition rate and the limitation of the current technology[3]. For example, the dictation machine that is developed for non-specific speech recognition is only suitable for standard mandarin and it has poor adaptability to dialects, isolated words and spoken languages. It is difficult for the annual news system to retrieve speech key words because of the high cost of required hardware. But speech recognition products have not stopped the development and previous research has accumulated a foundation for later small-scale product development, especially for those household toy consumer products market with low recognition rate, there is a wide range of application prospect.

HMM algorithm model and principle

Hidden Markov Model (HMM) is a probabilistic model, which is presented with parameter and used to describe the statistical characteristics of stochastic processes. It is evolved from a Markov chain. It may be a one-dimensional observation sequence, a coded symbol sequence, or a multidimensional vector sequence[4]. The basic theory of HMM was developed in 1970s. It has not been successfully used in speech processing until the last ten or twenty years. It is a significant progress in the digital processing of speech signals in 1980s by using the model to describe the generation of speech signals. Great achievements have been made in solving the problem of speech recognition by using this method. The basic theory and various practical algorithms are important cornerstones of modern speech recognition. The definition of Hidden Markov Model is as following:

Definition 1: The observation sequence $O = o_1 o_2 \cdots o_r$ is a random sequence. A HMM with N statuses $(S_1, S_2, \cdots, S_N)$ is represented by three parameters group $\lambda = (\pi, A, B)$, which is used to describe a probabilistic model for statistical properties of random sequences. Among them,

1. $\pi = [\pi_1, \pi_2, \cdots, \pi_N]$ is the initial distribution, which is used to describe the probability distribution of states of observation sequence O in the model when $t = 1$ and the state of $\pi_i = P(q_i = S_t) \quad i = 1, 2, \cdots, N$ is q_i. That is, of course, it satisfies Formula 1.

$$\sum_{i=1}^{N} \pi_i = 1 \qquad (1)$$

2. $A = \{a_{ij} \,|\, i, j = 1, 2, \cdots, N\}$ is the state transition probability matrix. Only first-order HMM is considered here and the present status q_t is only related to the state of the previous moment q_{t-1}, so Formula 2 can be obtained.

$$\begin{aligned} a_{ij} &= P(q_t = s_t \,|\, q_{t-1} = s_i, q_{t-2} = s_k \cdots) \\ &= P(q_t = s_j \,|\, q_{t-1} = s_i) \end{aligned} \qquad (2)$$

The above formula satisfies the condition of Formula 3.

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad (3)$$

3. B is the observation sequence. O is the distribution of any observation in the observation probability space of each state. There are two types of this distribution: the discrete type and the continuous type, respectively corresponding to discrete HMM and continuous HMM. Their distributions are: 1) In the case of discrete HMM, the observation sequence is a symbolic sequence and B is a probability matrix(Referring to formula 4).

$$B = \{b_j(k),\ j = 1, 2, \cdots, N; k = 1, 2, \cdots, M\} \quad (4)$$

The above formula satisfies the condition of Formula 5.

$$\sum_{k=1}^{M} b_j(k) = 1 \quad (5)$$

M is the total number of concentrated symbols used to encode symbols and When vector is used to make quantization on coding, M is the size of code book and j is the sequence number of status. 2). In the case of continuous HMM, the observation sequence is a vector quantization sequence and B is the collection of N probability density functions with D dimension. Please refer to Formula 6.

$$B = \{b_j(O),\ j = 1, 2, \cdots, N\} \quad (6)$$

O is any vector of observation vector space and each density function satisfies the condition of normalization. Please refer to Formula 7.

$$\int_{\Omega_j} b_j(o)do = 1 \quad (7)$$

Ω_j presents the observation probability space of status j, which can be all the space of vector O and can also be a subspace or an area of it. The above is the complete definition and description of Hidden Markov Model. From this definition, it can be seen that HMM uses the initial distribution and the state transition probability matrix to describe the statistical properties of a finite sequence of random sequences as the first order Markov chains of finite states. But unlike Markov chains, which allows the state of the current state to be known each observation, but by each observation, it can only estimate the probability of the current state. That is to say, it has double randomness and is a double stochastic process.

The application of HMM in speech recognition

The signal is a physical process that can be discrete, such as letters in a finite alphabet and code words in a code book, etc.;It can also be non-stationary, that is, the nature of the signal varies at any moment.

HMM uses Markov chains to simulate changes in the statistical properties of signals, which are indirectly described by observation sequences. Therefore, it is a double stochastic process. A speech signal itself is an observable sequence and it is the parameter flow of phoneme that is not observable in the brain and is issued on the basis of speech needs and grammatical knowledge, therefore, the precise model of speech signals must be described by Hidden Markov Models. The speech signal changes over time, which indicates the uncertainty of the speech signal. In order to describe the characteristics of speech signals over time, the concept of “state” is adopted, and the change of speech feature is presented by transferring from one state to another. By using HMM technology, only the system with limited states can be taken as speech generation model. Each state produces a finite output. In the generation of a word, the system continuously transfers from one state to another, circles to another state, each state produces an output until that the entire word output is completed, and an example of this model is shown in Figure 1.

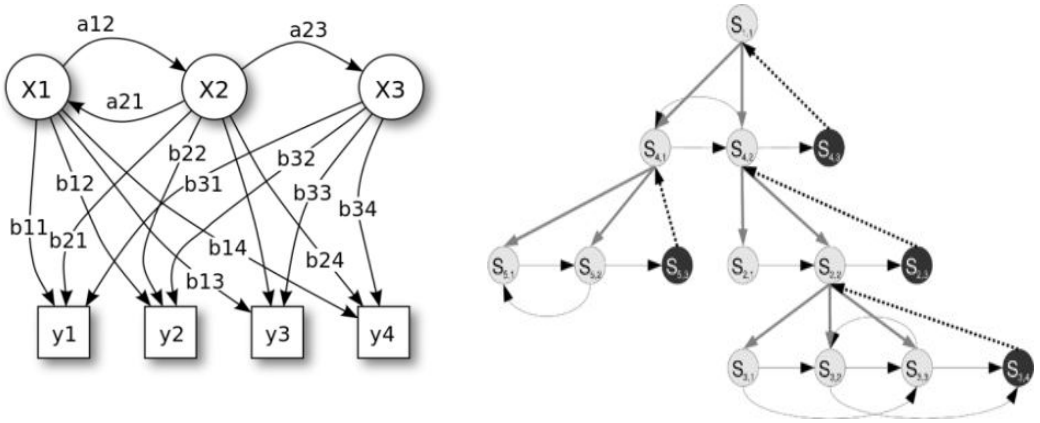

Figure 1 Shapes of two typical HMM Markov chains

Design of speech recognition system

According to the various characteristic parameters mentioned above, a set of optimal parameters is selected to form a “feature vector”. The voice of a particular person that has been identified is stored in vector memory in a vector manner and this signal is subjected to “signal pre-processing” first, according to its energy and zero crossing rate change, a simple real-time endpoint detection is done so as to remove the mute to obtain the time domain of the input speech and conduct spectral feature extraction based on this. Then, according to the input speech spectrum, results are analyzed and the energy distribution of each segment is calculated, the light consonants, voiced consonants and vowels are identified. After determining the vowel, voiced segment, please extend the front and back ends to search for frames that contain speech endpoints. Now the feature vector is the characteristic vector to be recognized. Comparing this set of features with the corresponding features of the pre- stored standard template, the distance between them can be obtained, and then the result is sent to “the decision logic section” to make a decision.

Because even if the same person has the same word, there will be some changes in the speed, so the distance from the beginning to the end of the word will be different, and the duration of the syllable will be different. Therefore, first of all, after the “amplitude detection” and “zero pass rate detection” part, the result will be sent to the beginning-to-end-point detection and the vibration discrimination part as well as the beginning and end part of each word, thus voice can be divided into words. Then, after a time corresponding step, there is the conduct of adjusting the speed of incoming signals at each time within the allowable range and then comparing to the standard signal.

In this step, we obtain a corresponding relationship between “the time axis of the incoming signal” and “the time axis of the standard signal”. In such relationship, The length of each short-term segment should be divided in detecting the characteristics of the signal and the length information can be provided to “pitch detection” or “predictive coefficient detection”. In feature detection section, based on the information, the former part will break the incoming voice into short periods, calculate each characteristic and then reverse-send the counted features to “time response procedure and distance calculation” part so as to work out the distance between each feature.

Conclusion

It can be predicted that speech recognition system will be used more widely in recent few decades and a wide range of language system products will appear on the market. But in the short term, speech recognition systems that are comparable to those of people can not be created and building such a system is still a great challenge. We can only move step by step towards improving the voice recognition system, however, it' hard to predict when we can build a speech recognition system that is as perfect as a human. As the development of Large Scale Integration technology has a huge impact on our society today, speech recognition technology will have a wide range of applications.

References

[1] Paiva Proenca, Kseniya, Kris Demuynck, and Dirk Van Compernolle. "Designing syllable models for an HMM based speech recognition system." Lecture notes in artificial intelligence. Vol. 9811. 2016.

[2] A-li H U I, ZHOU Q. State Identification of Insulators Flashover via Multi-Mode HMM Based on Multi-Features of Frequency Spectrum[J]. Journal of Applied Science and Engineering Innovation, 2016, 3(6): 202-208.

[3] Chen P Y, Pai N S, Chen G Y, et al. Design and implementation of a speech controlled omnidirectional mobile robot using a DTW-based recognition algorithm[C]//Applied System Innovation: Proceedings of the 2015 International Conference on Applied System Innovation (ICASI 2015), May 22-27, 2015, Osaka, Japan. CRC Press, 2016: 275.

[4] Jun N, Dongbo Z, Xu Z. Education Big Data Information Security Policy Analysis and Research in Cloud Computing Environment[J]. Journal of Applied Science and Engineering Innovation, 2016, 2(1): 157-160.