Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 3 (2012) 583 – 588

AASRI Procedia

www.elsevier.com/locate/procedia

2012 AASRI Conference on Modeling, Identification and Control

Human-machine Interaction Based on Voice

Jia-ni Lu[a], Hua Qian[a,]*, Ai-ping Xiao[a], Miao-wen Shi[b]

[a]*School of Technology, Beijing Forestry University, Beijing 100083, China*
[b]*School of Software Engineering,*
Beijing University of Posts and Telecommunications, Beijing 102209, China

Abstract

The human-machine system holds the important position in every field of man and machine. It is also a key technology for man to operate one machine. Voice-based identification is one of popular way to contact the user and device. It has been used widely in many areas with the development of hardware technologies and the improvement of mathematical algorithms. The article here elaborates the theory and algorithms which used in the system, at the same time shows the results of experiment.

Keywords: Man-machine interaction; Voice recognition; Pretreatment; Feature extraction; Algorithms.

1. Introduction

The definitions of man-machine interaction are extensive. Simply stated, it is a very important technology for man to operate one machine. With all kinds of devices, we can see that the variety of interactions must be rich. In 2008, Bill Gates, the CEO of the Microsoft Corporation, puts forward the concept of "natural user interface", and predicts the mode of human computer interaction will be changed a lot in the next few years. He told the reporters the keyboard and the mouse will replaced by touch screen, voice recognition, machine vision. Nowadays more simple and more convenient, but very effective control method becomes a trend. A

* Corresponding author: Hua Qian. Tel.: +1-365-137-0769.
E-mail address: qianhua@bjfu.edu.cn.

doi:10.1016/j.aasri.2012.11.092

series of artificial intelligence methods, including voice recognition and machine vision has been applied to reality. This paper reports a study on the key technologies of voice recognition.

2. Algorithm and Realization

Speech recognition uses voice signals as input instead of traditional keyboard, mouse and touch screen. It is more convenient and faster. It is also a kind of technology makes the machine to adapt to people. With the development of hardware technologies and the improvement of mathematical algorithms, it becomes more and more consummate.

The speech recognition is a kind of pattern recognition essentially. Feature extraction, pattern matching and reference library are the basic units of it. The process of the system is shown as figure 1[1].

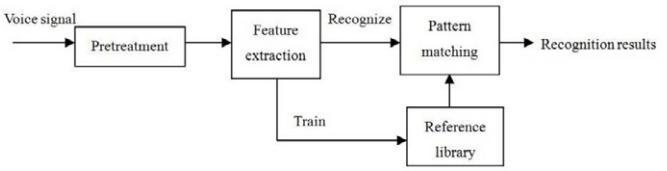

Fig. 1. Process of the voice recognition

2.1. Pretreatment

Before recognition, we must do some pretreatments for voice signal firstly which include pre-emphasis, frames, windowed and voice activity detection. Each step has a great impact on the performance of the entire system.

Functions of pre-emphasis are improve the high frequency part of signal. Because of the energy of the speech signal attenuates when the frequency increases, we can use pre-emphasis to compensate the high frequency part of signal which depressed by vocal organs. The formula named "first order, zeroes, digital filter" [2] is used most frequently.

$$H(z) = 1 - bz^{-1}, \qquad b = [0.90, 0.98] \tag{1}$$

The "b" is called the pre-emphasis coefficient. Figure 2 shows the oscillogram of "tell a joke" before and after Pre-emphasis.

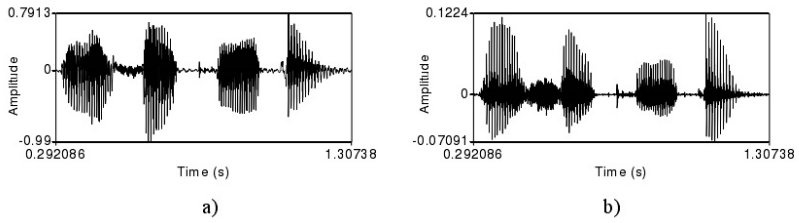

Fig. 2. The oscillogram of "tell a joke" before and after Pre-emphasis

In fact we divide the voice signals into some time fragments (about 10~30 ms), called "frame" when we deal with it. The frame is the smallest unit of signal processing. Between the frame and the adjacent frame, there is a certain overlap, generally overlapped by 1/3 or 1/ 2. The non overlapping part is called the frame shift. In general, frame-length takes 25ms when frame shift takes 10ms. And then we should windowed the signal in order to make the main lobe obviously and minor lobe less obviously. The reason for doing this is to eliminate the discontinuity of speech frame at the left and right. Hamming window is the most commonly used window function[3].

$$W(n)=\begin{cases} 0.54-0.46\cos(\dfrac{2\pi n}{N-1}) & 0<n<N-1 \\ 0 & n=other \end{cases} \tag{2}$$

Figure 3 shows the oscillogram of "tell a joke" 25ms abstraction before and after adding a hatch.

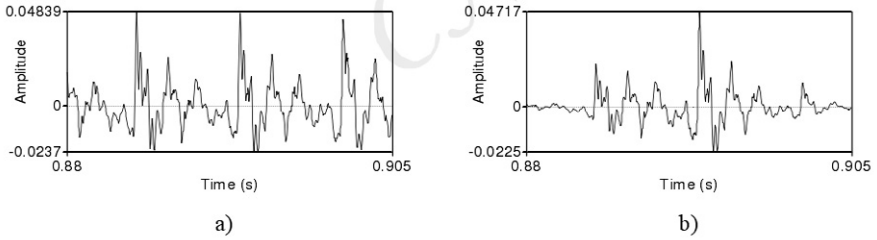

Fig. 3. The oscillogram of "tell a joke" 25ms abstraction before and after adding a hatch

Voice activity detection is a critical step in the pretreatment of the signal, only if the device determines the endpoint of the signal correctly can it deal with the voice by rule and line. We choose dual threshold method for voice activity detection. There are differences between two terms: short-term energy analysis and short-term zero rate algorithm analysis. For the n frame signal, $x_n(m)$, the formula of short-term energy is as follow:

$$E_n=\sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2 \tag{3}$$

And short-term zero rate algorithm is look this:

$$Z_n=\sum_{m=-\infty}^{\infty}|\operatorname{sgn}[x(m)]-\operatorname{sgn}[x(m-1)]|\,w(n-m) \tag{4}$$

Sgn means a signum expressed as function:

$$W(n)=\begin{cases} 1 & x(n)\geq 0 \\ -1 & x(n)<0 \end{cases} \tag{5}$$

By using the dual threshold method for voice activity detection, we should set two thresholds for short-term energy analysis and short-term zero rate algorithm analysis, one high and one low for each. The low one is sensitive for signal changes so it is easy to be overstepped, but the high one only can be overstepped by the signal which reaches to some intensity. The method is divided into four stages: mute, transition, voice and end. During the mute, if one of the energy or the zero rate over the low threshold, the flag will be changed. This indicates the process enters the transition, but now we cannot be sure whether there really is a voice signal or not, therefore if the numerical value down to below the low threshold, the current state goes back to the mute. But if any one of the parameter oversteps the high threshold, we can determine that it get into the voice part. At this part if tow parameters down to below the threshold at the same time the time is shorter than the time threshold we believe that this speech signal is noise. If not we label the endpoints and return.

2.2. Feature extraction

We choose the Mel Frequency Cepstrum Coeficients(MFCC) to do the feature extraction from many kinds of methods. The theory is based on linear predictive coding (LPC), according to phonation mechanism, and psychology model of auditory system. Mel frequency can be formulated as follow:

$$f_{mel}=2595\log(1+\frac{f}{700}) \tag{6}$$

The MFCC conisits by four steps:

- Fast Fourier Transform (FFT). The characteristics of the voice signal reflected by the energy distribution in the frequency domain. After each frame of speech signal is multiplied by a Hamming window, it also needs to pass the FFT to get the energy distribution in the frequency domain.
- Triangular band-pass filter. Let's assume the number of overlapping triangular band-pass filter is M. Then we can figure out the Kth output of the filter is x (k). Diagrammatic sketch of triangle band-pass filter is shown as Fig 4.

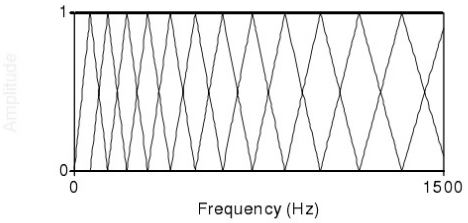

Fig. 4. Diagrammatic sketch of triangle band-pass filter.

- Inverse discrete cosine transform. Firstly take the log of the output of each filter to get power spectrum, then use the inverse discrete cosine transform in order to get several MFCC coefficients, finally take about 12~16 coefficients.

- The MFCC features will be regard as static characteristics, and then do the first-order and second-order differential on them to summarize the dynamic characteristics.

2.3. Recognition Algorithms

We take the Hidden Markov Models (HMM) to achieve a non-specific recognition system. HMM is a probability model for the statistical properties of the random process evolved by the Markov chain. A complete HMM needs to be given some parameters are as follows: state number N, the number of observation symbols M and the aggregation, as well as the probability distribution A, B, p, where A is the state transition probability matrix; B is the observation symbol probability matrix; p is the initial state probabilities. Typically, the following formulas are here to represent a particular model.

$$l = (p, A, B) \tag{7}$$

The HMM can be thought of as the generator of one observation sequence(observation sequence Q), where T is the length. Q expressions are as follow:

$$Q = (Q_1, Q_2, Q_3 \ldots Q_T) \tag{8}$$

For HMM model, there are three basic questions need to be addressed:

- Calculate the output probability. Given the observed sequence Q and the model l, how to calculate the output probability $p(Q \mid l)$ quickly and efficiently.
- Find the best state sequence. Now Q and l are known, how to find out the best state sequence.
- Model training. How to adjust the model l such that the conditional probability $p(Q \mid l)$ is the maximum.

Consider one problem of a non-specific and N-word identification: For each word in the word table, we need to design an X-state HMM. The speech signal of one given word as an observation sequence. Here, it is assumed that the codebooks with the M codebooks vector encoding for speech spectrum vectors, each observation is an index, and the index corresponding to the codebook vectors is closest to the spectrum of the speech signal vector in a way. The voice recognition system is to draw recognition results depend on this index.

3. The experimental results

In order to improve the credibility of the results of experiments, we tested the system in tow different environment, slightly noisy and noisy, the personnel involved in the testing were strict screening. In both environments there were 20 adults (10 males, 10 females) and 20 children (10 male, 10 female) took the test, everyone for 100 words. The test results shown in Table 1:

Table 1. The experimental results

	Detectable rate	False detectable rate
Slightly noisy	89%	6%
Noisy	83%	13%
Adults	90%	5%

Children	85%	10%

From the experimental results, we can know that the environment has a very important impact on the recognition results. In addition, the accuracy of the pronunciation also affect the recognition rate. The slightly noisy environment has a higher detectable rate. The pronunciation of adults is more accurate and more stable than children, so their recognition rate is also higher than them.

4. Summary

This paper describes a voice-based interactive system, and the main part of the speech recognition, including pretreatment, feature extraction, and recognition algorithms are explained in detail and tested by some experiments. Speech recognition systems have become increasingly common to our life, it is bound to bring us a new interactive experience.

Acknowledgements

This subject is funded by state forestry administration 948 item.

This full name of this foundation item is state forestry administration introduced advanced forestry science and technology project "Critical technology introduction of Flotation Column recycling fiber from deinking pulp":Number:2009-4-57.

References

[1] Zhao Li. The basis of robotics. M Beijing: China Machine Press 2009; 2-5.
[2] Bai Sun-xian. Research of Chinese isolated word voice recognition. D Chengdu: Southwest Jiaotong University; 2009; 15-16.
[3] Shen Hong-yu. Research and Implementation of the DSP-based voice system. D Jiangsu: Jiangnan university; 2008; 6-7.
[4] Lu Qing-qi, Bai Yan-yan. Two judgments of the dual-threshold-based speech endpoint detection method. J Electronic Science and Technology. 2012(25); 13-15.