國立臺灣大學工程科學及海洋工程學研究所

碩士論文

Department of Engineering Science and Ocean Engineering

College of Engineering

National Taiwan University

Master Thesis

使用寵物型機器人改善朗讀節奏之應用

Improving Oral Reading Rhythm by a Pet-like Robot

劉彥圻

Yen-Chi Liu

指導教授：郭振華 博士

Advisor: Jen-Hwa Guo, Ph.D.

中華民國 99 年 9 月

September, 2010

致謝

首先要感謝指導教授郭振華老師細心的指導，並給予課業及生活上的建議及幫助，並且提供良好的學習環境，讓我得以順利完成學業。十分感謝口試委員江茂雄老師、王傑智老師和洪儷瑜老師給予論文改進的意見與指正，使的本文更加完整和充實。

感謝水下實驗室裡的學長以及同學：冠宇、惟果、思亮、瑞閔、佳縉、譽元、寶麒、盛煒、銓熙、偉翰、柏諠、凌吉、宗穎在這幾年來提供的幫助。還有在最後這段時間一起奮鬥的夥伴玉龍、定橋、明毅、信豪，尤其是信豪在實驗軟硬體上的大力支持，我會永遠記得這段共患難的日子。還有奕倫以及丹林，祝你們在學術研究上能出類拔萃，順利拿到博士學位。同時也謝謝實驗室的學弟們：柏昇、傳迪、苗鋒、家毅、瑞軒，有你們的幫助和所帶來的歡笑，給我很大的幫助和鼓舞。

此外，感謝運動與控制實驗室意明在校時間的陪伴，還有應用聲學實驗室文馨在錄音上的協助。特別感謝好友芷庭，在英文寫作上給予的建議和幫助。

最後要感謝的是我的父母及家人，提供物質及心靈上的後盾，還有女友古莉這段時間的不離不棄，是我完成論文的最大動力。

由衷感謝這幾年來給予幫助的所有人，使得本文得以順利完成。

摘要

本研究目的為使用寵物型機器人，在和使用者的互動朗讀過程中，評估閱讀者文句節奏狀況，並給予回饋進而幫助閱讀流暢度的改善。辨識朗讀者節奏，必須先利用語音端點偵測技術，找到字的端點。本研究機器人的端點偵測系統，是使用過零率、能量、頻譜、自相關函數等特徵，訓練隱藏式馬可夫模型之分類器。互動的機制上，在機器人上使用兩個垂直方向的馬達，發展出具節拍器功能之尾巴，用以給予使用者閱讀節奏之引導。

為了評估讀者和機器人相互影響關係以及讀者節奏狀態。以句子為單位，定義字相位模型，並用此定義同步參數以及節奏參數。其中同步參數用以評估讀者和節拍器的相互關係，節奏參數用以定義句子結構以及評估讀者節奏。評估使用者節奏參數，若其沒有達到指定的朗讀節奏，在下一次的朗讀過程，節拍器回授使用者一個更顯著差異的時間差，使之更具朗讀節奏。而時間的改變量也將根據同步參數不同而有所調整。

本研究實驗部份，首先藉由分析有節拍器引導與否，顯示尾巴引導機制的效果。接著藉由重複的引導過程，觀察使用者朗讀狀態的改變。最後用隨機決定的朗讀文句，觀察使用者受節拍引導的影響。未來，可用以陪伴幼童讀書以及幫助其提升對文章節奏的熟悉度。

關鍵詞:朗讀,節奏,控制,隱藏式馬可夫模型,機器人

Abstract

In this study, the pet-like dog that interacts with reader for enhancing the reading cadenced is being developed. Acoustic behaviors of voice signals are recorded to analyze the condition during interaction. The reader reading rhythm is found by endpoint detector. The endpoint detector is a classifier and based on the Hidden Markov Model. It is trained by nonverbal auditory cues, such as zero-crossing rate, energy, spectrum, and autocorrelation. The spectrum and autocorrelation are chosen to recognize the periodic signal.

To provide guidance of oral reading, the robot's tail is developed into a two directions metronome by the two perpendicular motors. The goal is to classify reading states of the user, so the phase model of the words' period in a sentence is defined, then the synchronization and rhythm parameter are defined by word phases. The synchronization parameter characterizes the users' response models with respect to the tail input. And the rhythm of the sentence is characterized by the rhythm parameter. Measuring the rhythm parameter, if the rhythm parameter is far from the commanded rhythm, a new pace is then set up to control the metronome for the sentence reading guidance. The guidance rule for a specific user considers not only the rhythm parameter difference but also the synchronization parameter of the user.

Experiments were conducted to demonstrate the effect of guidance on oral reading performance. Rhythm parameters being controlled to approach better reading fluency is observed under the proposed guidance rule.

Keywords: oral reading, rhythm, control, hidden Markov model, robot

Table of Content

List of Figure

List of Symbol

$x(t)$	Continuous sound signal
$x_p(t)$	Signal after sampling
$p(t)$	Impulse train of time
$x(n)$	Discrete signal
T_s	Sampling period
F_s	Sampling frequency
N_F	Frame duration
M_F	Frame overlap
L	Number of total frames
l	Frame index
$f_l(n)$	Signal after frame blocking
$w(n)$	Hamming window
$x_t(n)$	Signal after Hamming window
$E(l)$	Short-time log energy
$Z(l)$	Zero crossing rate
$a(k)$	Short-time normalized autocorrelation of the signal
$A_v(l)$	Short-time normalized autocorrelation maximum peak value

$A_p(l)$	Short-time normalized autocorrelation number of peaks
$X(k)$	Fast Fourier transform of the signal
$F_v(l)$	Fast Fourier transform maximum peak value
$F_f(l)$	Fast Fourier transform maximum peak value frequency
$H_s(l)$	Spectral entropy
q_t	State of HMM model
N	The number of states in the HMM model
M	The number of distinct observation symbols per state
S_i	The i_{th} state in the model
o_t	Observation of HMM model
V	Observation vector of HMM model
A	HMM state transition probability distribution
a_{ij}	The probability from state i to state j
B	Observation symbol probability distribution
b_j	The probability of the observed variable in state j
π	The initial state distribution
π_i	The initial probability that at state i
λ	The complete parameter set of an HMM

O	The observation sequence
T	Sequence length
$\alpha_t(i)$	Forward probability
$\beta_t(i)$	backward probability
$\xi_t(i,j)$	Probability of being in state S_i at time t and state S_j at time $t+1$
$\gamma_t(i)$	the probability of being in state S_i at time t
$\delta_t(i)$	Maximum probability along the best probable state sequence path of a given observation sequence after t instants and being in state i
$\psi_t(i)$	Backtracked function by best state sequence
Q	Best state sequence
$S_v(l)$	State of the voice
t_j	Teacher’ s reading time
t_k	Reader’ s reading time
N_w	The number of words in the sentence
T_s	Total time of the sentence
θ_j	Word phases of teacher

θ_k	Word phases of reader
ω	Reading speed parameter
p_θ	Synchronization parameter
D	Rhythm parameter
D_1	Level 1 of the rhythm parameter
D_2	Level 2 of the rhythm parameter
D_3	Level 3 of the rhythm parameter
D'	The new produced rhythm parameter
D_m	The measured rhythm parameter
p_m	Maximum synchronization error
p_w	Word synchronization error
C_2	The ratio of rhythm parameter
C_3	The ratio of rhythm parameter

Chapter 1 Introduction

1.1 Motivation

A robotic pet dog is designed to enhance children's love of reading by interacting with children. A robotic pet dog for reading assistance is considered to be effective in drawing children attention and creating intimacy. Turn-taking interaction is a general way to help children learning by parents or teachers. Traditionally, there are some assisting means employed as the educational guidance, for example, gesture or handclap. The tail mechanism of the robotic dog was developed to implement this function. Acoustic behaviors of voice signals are recorded while the child reads as a mean to analyze the condition of interaction. The goal of the study is to use the robotic pet dog as a tool to classify reading states of the child, and furthermore try to change the state into a good-learning condition, particularly for the oral reading fluency improvement. If the child is not cadenced of the sentence, this study demonstrated that the reading fluency can be improved by giving the more significant tempo provides by the robotic pet dog. The robot can play as an assistant or even a teacher in the interaction.

1.2 Literature Review

Kuhn and Stahl provide instructional strategies that are effective in promoting

fluency among beginning readers in [1]. It also gives a review of many studies that have attempted to improve fluency. It indicated that assisted approaches, such as reading-while-listening, seem to be more effective than non-assisted approaches, such as repeated reading. The basic ideal of the robot instructional strategies was derived from this literature.

The studies by Mastropieri, Leinart and Scruggs [2] reviewed the strategies for increasing reading fluency such as repeated reading, reading with computer or a reading with a peer. They indicated that careful application of combing some of these strategies can meaningfully improve reading fluency. The study was inspired by their work to design a robot for accompany readers.

Rasinski [3] referred to reading fluency as the reader's ability to develop control over surface-level text processing, so that he or she can focus on understanding the deeper levels of meaning embedded in the text. And he also referred to reading fluency has three important dimension that build a bridge to comprehension, such that accuracy in word decoding, automatic processing, and prosodic reading. In our work, our goal is to make sure that the reader does not place the equal emphasis on every word, which is related to the definition of the prosodic reading.

Then we are focus on the rhythm. Wolff has the discussion in [4] that dyslexic students contrast to normal readers. The experiment is which had inordinate difficulty

reproducing simple motor rhythms by finger tapping and similar difficulty reproducing the appropriate speech rhythm of linguistically neutral nonsense syllables. Then the dyslexic students took significantly longer than normal readers did to recalibrate their tapping cadence and switch back to the anticipation mode, after an abrupt change in the metronome rate did. It provides the application of changing reading rhythm by the metronome and relates to design the control equation of the metronome.

Brady [5] provides the study on the metronome effect on stuttering. The adult stutterers served as subjects in four experiments on the mechanism by which pacing speech with a metronome increases fluency. There is the marked increase in fluency experienced by most severe stutterers when they pace their speech with a metronome. Then the tail mechanism was designed as a metronome to guide the user reading.

Topping and Lindsay [6] synthesizes and analyses the research on the technique for non-professional tutoring of reading known as paired reading. The flow chart of the pair reading is used in our turn-taking reading procedure.

The literature [7] is an educational book about mandarin oral reading method for the young children. They use some simple symbols represent the different levels pause duration of the sentence. The rhythm parameter has defined in three levels just like the approach we adopted in this study.

In this work, endpoint detector was base on the hidden Markov model (HMM),

which was developed by Princeton (1980) [8]. This statistical model was usually used in speech recognition, but also used for the endpoint detecting in [9-11].

1.3 Thesis Organization

This thesis is divided into six chapters. In chapter 2, the signal processing techniques that were utilized to exact features for the endpoint detection. The endpoint detector based on a Hidden Markov Model was used to train and find the user state of reading are presented in chapter 3. Then the pet dog platform and the turn-taking reading model are introduced in chapter 4. In chapter 5, the application of this work and the discussion are presented. Finally, chapter 6 provides conclusions and future work.

Chapter 2 Voice Signal Processing

2.1 Introduction

This chapter shows the basic methods of voice signal processing. Oral reading sound is a continuously acoustic pressure wave signal. This chapter shows how to deal with those signals before the voice detector. First, the continuous acoustic signals turned into discrete signal by sampling, then block into frames. Second, some features of sounds can be exacted, for example, short time stationary energy, noisy autocorrelation are the features in time domain. And spectral entropy, pitch are features in frequency domain. Figure 2.1 is the flow chart of this chapter.

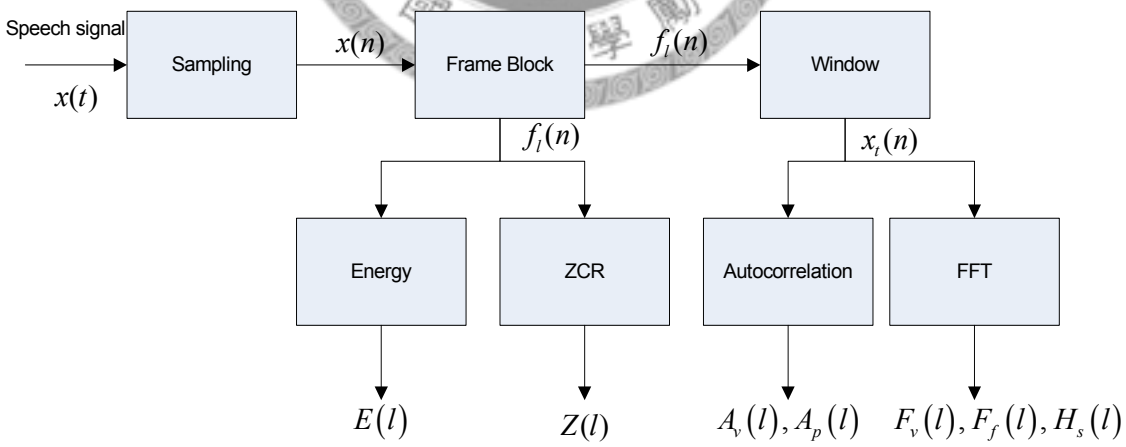

Fig. 2.1 Block diagram of the signal processing

2.2 Basic Voice Signal Processing

2.2.1 Sampling

Computer can only handle discrete signal, however, real world signals are continuous in time. Therefore, we have to convert continuous time signals to discrete signals, and this process is called sampling.

Mathematically, we perform sampling by multiplying a continuous signal by an impulse train which consists of periodic unit strength delta function in the domain of interest. Denoting the continuous waveform signal by $x(t)$, and $x_p(t)$ is the signal after sampling.

$$x_p(t) = x(t)p(t) \tag{2-1}$$

$p(t)$ is the impulse train, as follow:

$$p(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \tag{2-2}$$

The signal after sampling $x_p(t)$ has value only at $t = nT_s$, as the follow equation:

$$x_p(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_s)$$

$$= \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s) \tag{2-3}$$

So we can define a discrete time signal $x(n)$ such that:

$$x(n) = x(nT_s) \tag{2-4}$$

The sampling period of this discrete time signal is T_s, and the sampling frequency define by $F_s = 1/T_s$. We use the sampling frequency 8k Hz.

2.2.2 Frame Blocking

To deal with the discrete-time signal $x(n)$, framing is used to divide the speech signal into several sections. Frame blocking is to collect several samples to become a frame. We assumed that each frame's feature is invariant and features are processed from the speech signal frame by frame. The speech feature parameters can be extracted from each frame. Therefore, the variation in the speech signal can be observed.

There are two factors affect frame blocking: frame duration and frame overlap. Frame duration is the length of the frame, and frame overlap is the length of overlaps of two neighbor frames. The frame shift is defined as the difference between the frame

duration and the overlap, see the Fig 2.2

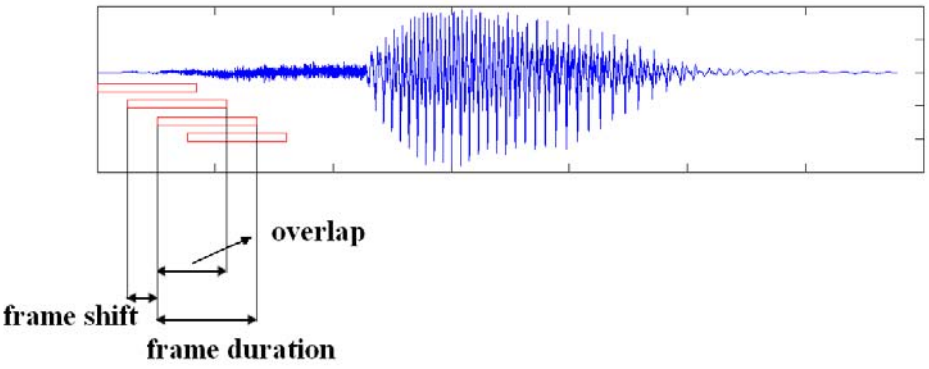

Fig. 2.2Frame Blocking

Longer duration come with fewer features could be detected in the speech signal. But if frame duration is small than a threshold, the result will be affected easily by the background noise. The feature extracted from each frame will be smoother by using frame overlap. Frame overlap is generally the half of frame duration.

The values of frame duration N_F and overlap M_F are 256 and 128 point in this paper, which correspond to 32 ms frames and separated by 16 ms when the sampling rate of the speech is 8k Hz.

The first frame consists of samples from 1 to N_F. The start sample of the second frame is $(N_F - M_F + 1)$ and the end sample is $(2N_F - M_F)$. We assume that there are L frames in the signal, and l is the number of frame. And the l frame is from $(l-1)(N_F - M_F)+1$ to $(l-1)(N_F - M_F) + N_F$. By applying the frame partitioning to $x(n)$, one will get L vectors of length N_F, and $f_l(n)$ denoted the l frame of the signal.

2.2.3 Windowing

In order to reduce the edge effect, we multiply a window. In this paper, the hamming window is chosen. The function of hamming window is defined as

$$w(n) = \begin{cases} 0.54 - 0.46 * \cos(\frac{2n\pi}{N_F - 1}), & 0 \le n \le N_F - 1 \\ 0 & , \ otherwise \end{cases} \tag{2.5}$$

The time and frequency responses of the Hamming window shows on Fig 2.3.

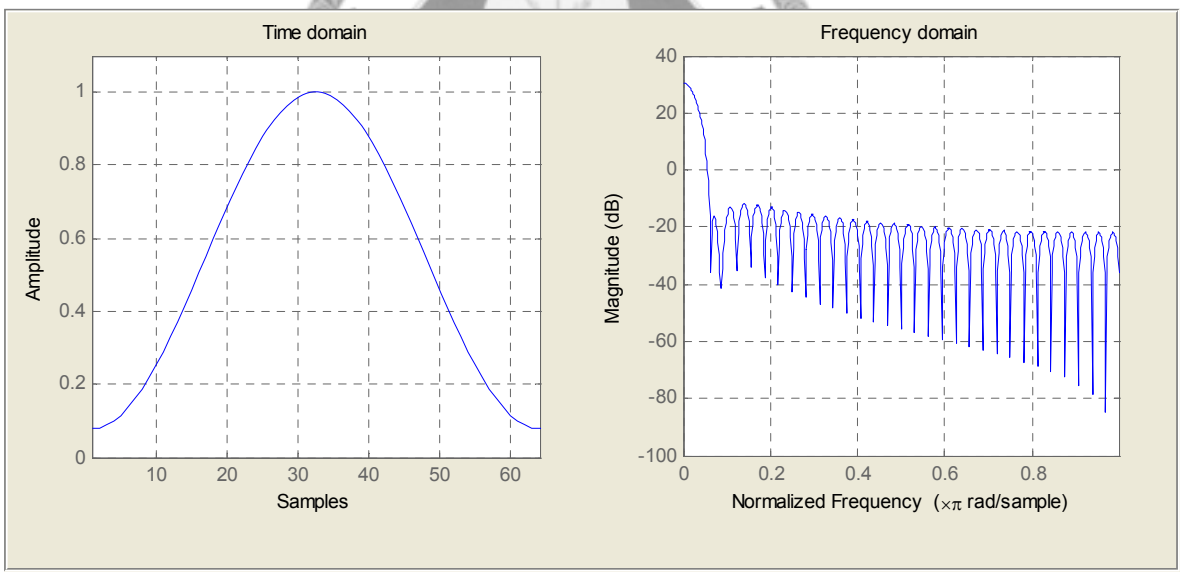

Fig. 2.3 Hamming window

By applying the $w(n)$ to every frame $f_l(n)$, then $x_t(n)$ is obtained.

2.3 Time Domain Feature

2.3.1 Energy

The short-time log energy was computed according to the following formula:

$$E(l) = \log\left[\sum_{n=l-N_F+1}^{l} f_l(n)^2\right] \tag{2.6}$$

Even though it is not very robust against noisy backgrounds and impulsive interferences, energy is still a fundamental component in many widely used endpoint detectors. The relationship between voice signal and energy shows on Fig. 2.4.

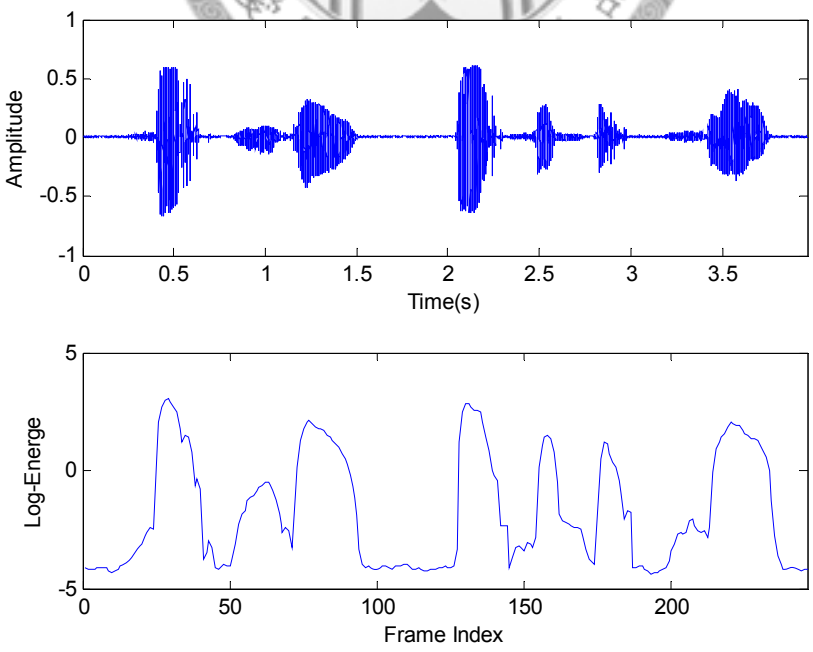

Fig. 2.4 Speech signal (up) and Log-Energy(down)

2.3.2 Zero-crossing rate(ZCR)

Zero-crossing rate (ZCR) is the number of zero point a wave passed in each frame. The ZCR can be mathematically defined as

$$Z(l) = \frac{1}{N_F} \sum_{n=l-N_F+1}^{l} 0.5 \left| \operatorname{sgn}\left[f_l(n) \right] - \operatorname{sgn}\left[f_l(n-1) \right] \right| \tag{2.7}$$

where the function $\operatorname{sgn}\left[f(n) \right]$ is defined as

$$\operatorname{sgn}\left[f(n) \right] = \begin{cases} 1 & , for\ f(n) \geq 0 \\ -1 & , otherwise \end{cases} \tag{2.8}$$

ZCR has an important feature that the ZCR of noise and aspiration signal are greater than that of speech and the ZCR curve is shows on Fig. 2.5.

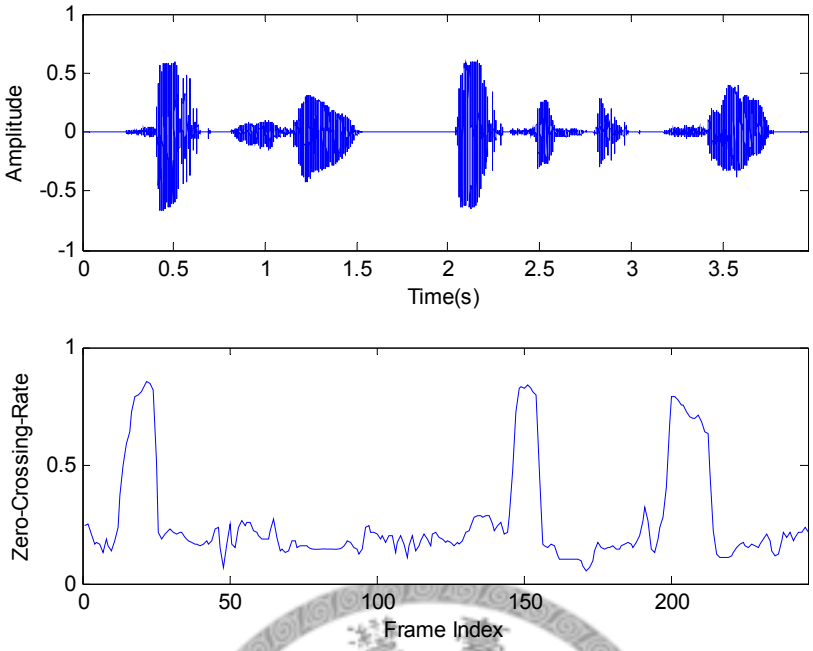

Fig. 2.5 Speech signal (up) and ZCR(down)

2.3.3 Autocorrelation

The short-time normalized autocorrelation of the signal $x_t(n)$ is defined as follow:

$$a(k)=\frac{\sum_{n=k}^{N_F} x_t(n)x_t(n-k)}{\left(\sum_{n=0}^{N_F-k} x_t(n)^2\right)^{\frac{1}{2}}\left(\sum_{n=k}^{N_F} x_t(n)^2\right)^{\frac{1}{2}}} \tag{2.9}$$

By the definition, $a(0)$ is guaranteed to be 1. It can be obtained a small number of strong peaks for voice frames because of their periodic component, on the other hand,

there is a large number of small peaks for unvoice frames. We can see its figure on the Fig. 2.6

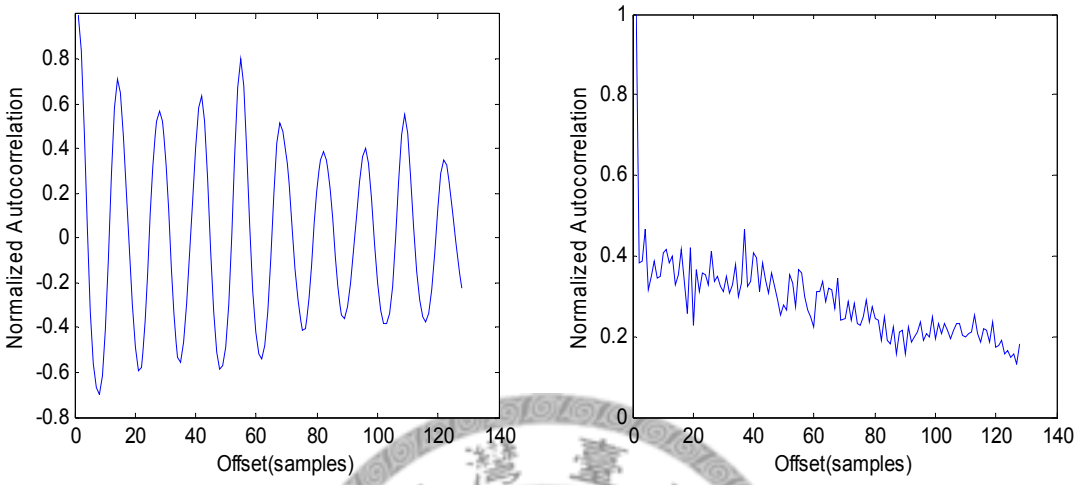

Fig. 2.6 Autocorrelation results for voiced (left) and an unvoiced (right) frane.

There is one significant problem to the standard normalized autocorrelation. The very small and noisy periodic signals will still result in strong peaks. To solve this problem, signals is to add a very low-power Gaussian noise signal to each frame before taking the autocorrelation. The figure 2.7 shows the relationship between the speech signal and the autocorrelogram.

We use the maximum peak value and the number of peaks as our two features. The maximum peak value and the number of peaks can be defined as $A_v(l)$ and $A_p(l)$

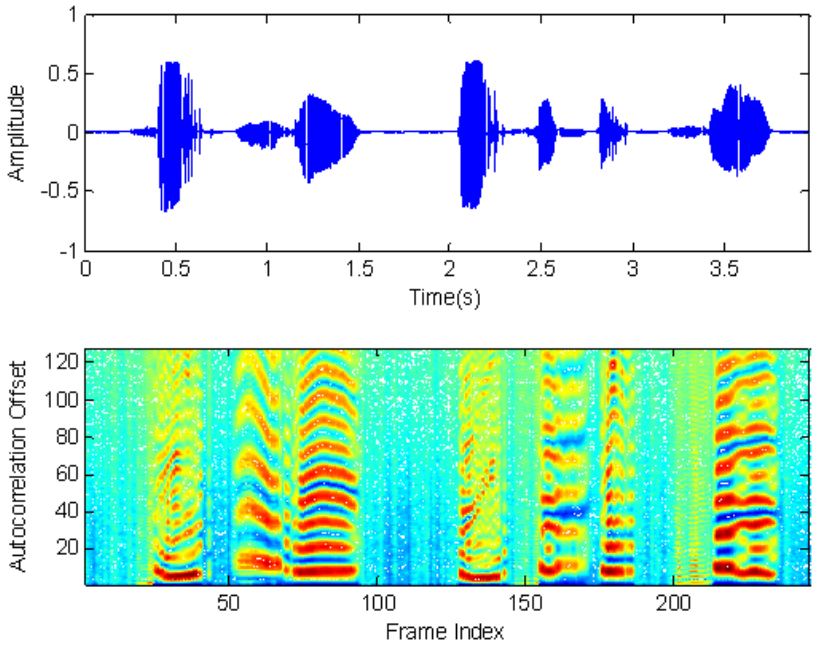

Fig. 2.7 Speech signal (up) and autocorrelogram(down)

2.4 Frequency Domain Feature

2.4.1 Fast Fourier Transform

The variation of signals in the time domain is hard to find out the characteristic of speaker signals. The signals transfer to frequency domain by FFT. Each frame after multiplying the hamming window change into frequency domain by following FFT equation:

$$X(k)=\sum_{n=0}^{N_F-1} x_t(n)e^{-j\frac{2\pi}{N_F}kn} \qquad (2.10)$$

$X(k)$ is a complex number, so it can be divided into two parts:

$$X(k) = |X(k)| \angle \phi_x(k) \tag{2.11}$$

$|X(k)|$ and k can show the spectrogram in Fig. 2.8.

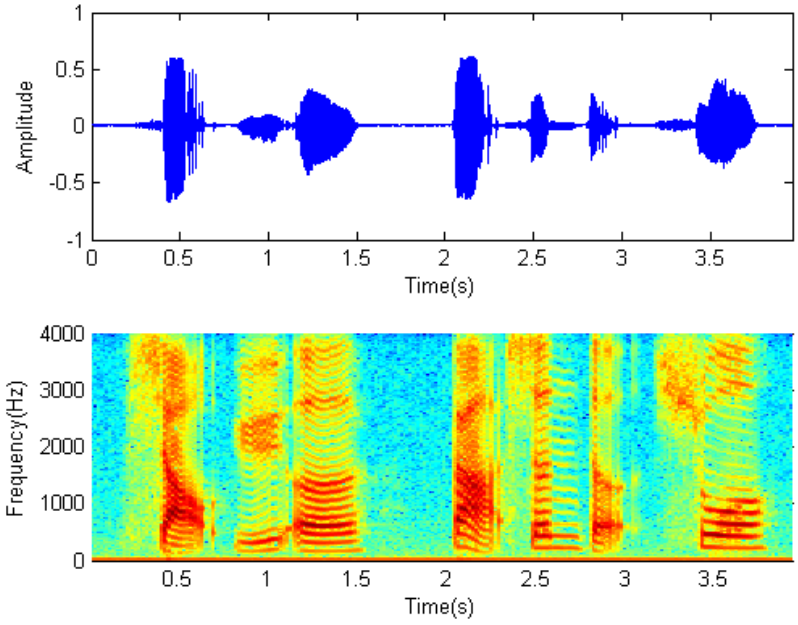

Fig. 2.8 Speech signal (up) and spectrogram (down)

We use the maximum peak value and the corresponding frequency as our two features. The maximum peak value and the number of peaks can be defined as $F_v(l)$ and $F_f(l)$

2.4.2 Spectral Entropy

Another key feature can be obtained as the FFT magnitudes. Voiced frames have a

series of very strong peaks resulting from the pitch period's Fourier transform. This result in the banded regions we have in the spectrograms and in a highly structured set of peaks as seen in the first panel of Fig. 2.9. In the unvoiced frames, as seen in the right panel, we see a fairly noisy spectrum, be it silence (with low magnitudes) or a plosive sound (higher magnitudes). We thus expect the entropy of a distribution taking this form to be relatively high.

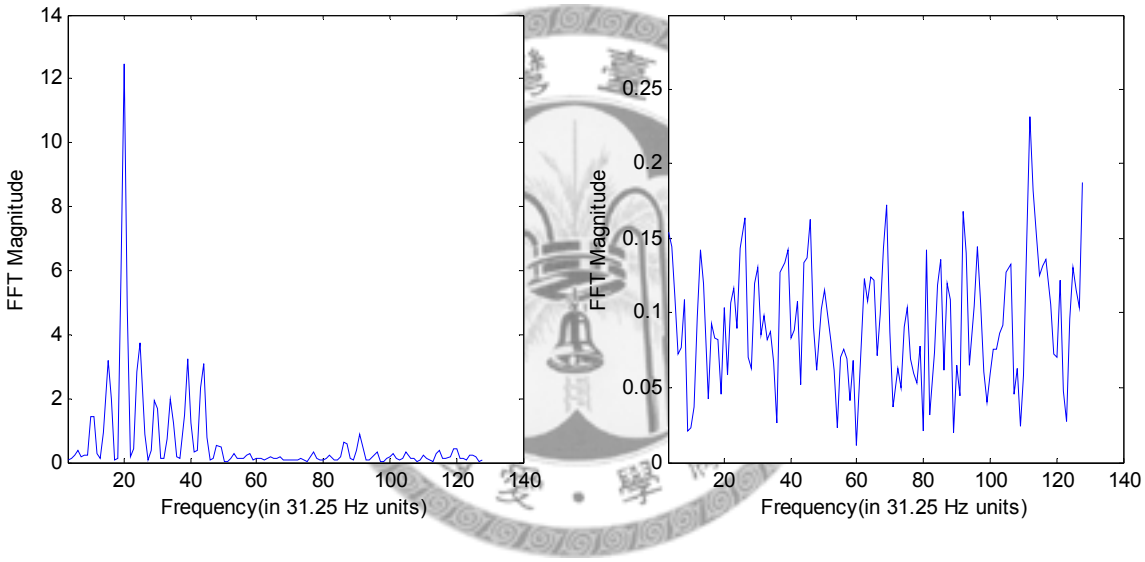

Fig. 2.9 FFT magnitude for a voiced (left) and an unvoiced (right) frame.

To compute the spectral entropy, $X(k)$ is normalized to make it into a proper distribution.

$$X_n(k) = \frac{X(k)}{\sum X(k)} \tag{2.12}$$

Normalizing in this way makes this feature invariant to the signal energy. The

entropy of the resulting distribution can be computed by following equation:

$$H_s = -\sum_k X_n(k) \log X_n(k) \tag{2.13}$$

In Fig 2.8, H_s is 3.76 for the voiced frame, and 4.72 for the right panel. The relationship between the signal and H_s will show in Fig. 2.10.

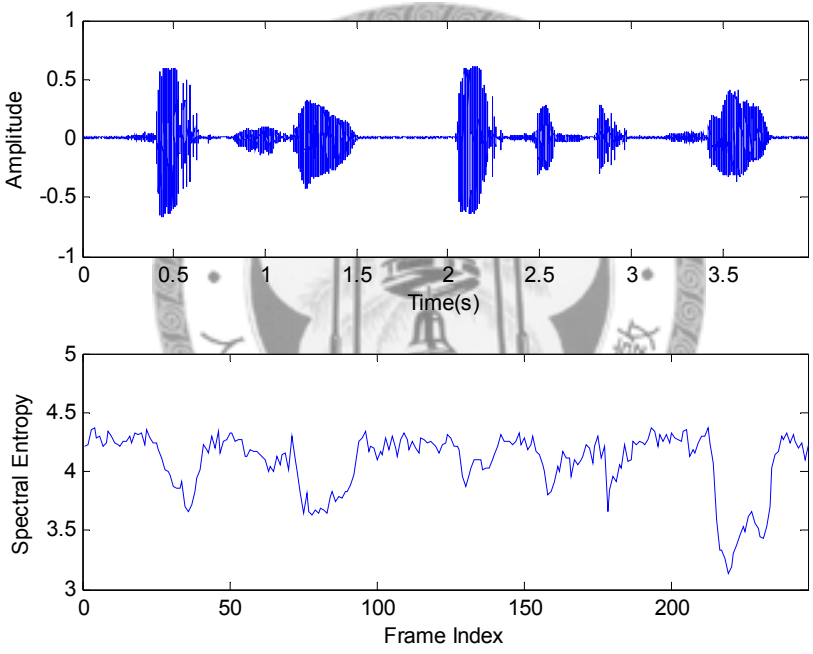

Fig. 2.10 Speech signal (up) and Spectral Entropy(down)

$H_s(l)$ is our final feature.

In this chapter, that there are six features be defined. They are short-time log energy $E(l)$, zero crossing rate $Z(l)$, autocorrelation maximum peak value $A_v(l)$, autocorrelation number of peaks $A_p(l)$, FFT maximum peak value $F_v(l)$, FFT

maximum peak value frequency $F_f(l)$, and spectral entropy $H_s(l)$.

These features are focus on recognizing voice and unvoice, and these features will be used in next chapter-- HMM-Based Endpoint Detector.

Chapter 3 HMM-Based Endpoint Detector

In this chapter we propose an effective, robust and computationally low-cost HMM-based start-endpoint detector. The features used for voice activity detection are energy, zero-crossing-rate, autocorrelation, and spectral entropy, which presented in chapter 2.

The endpoint module is a critical part in any spoken dialogue system. It cannot be recovered since missed speech fragments. Two main approaches are adopted in developing endpoint detectors: threshold based and classifier based.

The first class is the most widespread. The decision is performed according to one or more threshold. Its algorithms are generally simpler and faster to implement. Its major drawback consists in the need of careful tuning of many parameters, something that makes such algorithms sensitive to environmental variations. The second one uses the classifier substitutes the threshold. This method relies on general statistics rather than on local information.

In our application, it probably works in different environment, so it will be better to choose the second way to approach this goal. The classifier that we use is presented in the follow section.

3.1 Hidden Markov Models

To understand the problem of choosing the endpoint detection method, we must first examine the process of speech production and the Chinese language (We focus on Chinese language in this thesis). Speech can be broken up into two kinds of sounds: voiced and unvoiced. The voiced sounds are those that have a pitch, which we call vowel. The unvoiced sounds are everything else. Almost every Chinese word involves a vowel. Chinese is single syllable language. There are different states in the speech production process which we can't see. We introduce the HMM idea in the below.

We assume speech signal can be characterized as a parametric random process. We introduce the classifier hidden Markov model (HMM). The HMM model is often used in speech recognition, because the same reason of the pronounced state concept. But it is much simplicity in our application. We use the two states HMM model as a classifier. One represent the voice state and the other one represent the unvoice state. A hidden Markov model is a Markov model where the states q_t are not directly observable. Instead, we can observe another measurement o_t that is related to q_t by the stationary probability distribution $p(o_t \mid q_t)$, i.e., o_t is a probabilistic function of the unobserved state q_t .The graphical structure of the hidden Markov model is illustrated in Figure 3.1.

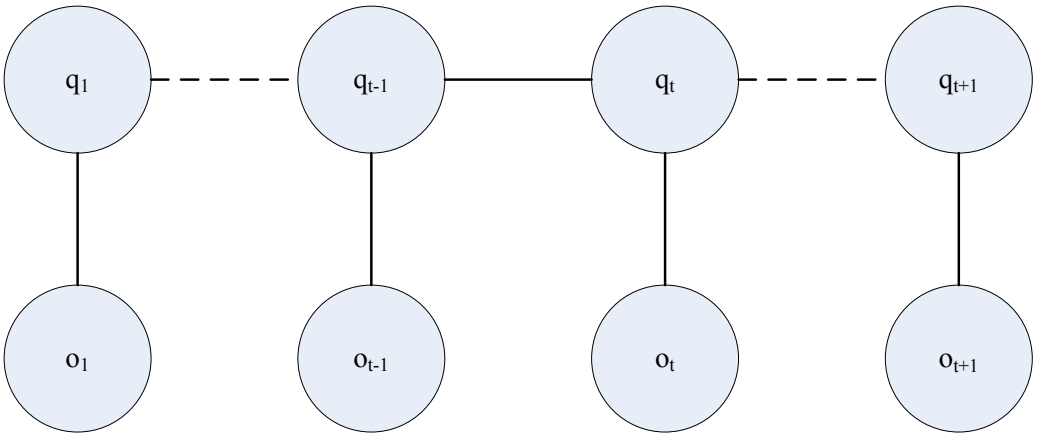

Fig 3.1 The graphical structure of a hidden Markov model

For a basic HMM, there some parameters that are necessary to described the process. An HMM is characterized by the following:

1. N, the number of states in the model. S_i, the i_{th} state in the model. ($N = 2$ in our application)

2. The observation vector:

$V = \{v_1, v_2, ..., v_M\}$, where M is the number of distinct observation symbols per state

3. The state transition probability distribution:

$A = \{a_{ij}\}$, a_{ij} is the probability from state i to state j, as the follow equation:

$$a_{ij} = P\left(q_{t+1} = S_j \mid q_t = S_i\right), \quad 1 \le i, j \le N \tag{3-1}$$

4. The observation symbol probability distribution in state j:

$B = \{b_j(v_k)\}$, $b_j(v_k)$ is the probability of the observed variable in state j, as the follow equation:

$$b_j(v_k) = P(o_t = v_k \mid q_t = S_j), \quad 1 \le j \le N, 1 \le k \le M \tag{3-2}$$

5. The initial state distribution:

$\pi = \{\pi_i\}$, π_i is the initial probability that at state i, as follow:

$$\pi_i = P(q_1 = S_i), \quad 1 \le i \le N \tag{3-3}$$

For convenience we usually use a compact notation $\lambda = (A, B, \pi)$ to indicate the complete parameter set of an HMM. It also requires specification of two model parameters (N and M).All parameters must be under the following constraints.

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N \tag{3-4}$$

$$\sum_{i}^{N} \pi_i = 1 \tag{3-5}$$

$$\sum_{j=1}^{M} b_j\left(v_k\right)=1 \tag{3-6}$$

An example of a HMM of a three-state Hidden Markov Model is shown as follows. There are three states, S_1, S_2, and S_3 in the model, and they generate *A, B*, and *C*, respectively. The state transition probability matrix is

$$A=\left\{a_{ij}\right\}=\begin{bmatrix}0.6 & 0.3 & 0.1\\ 0.1 & 0.7 & 0.2\\ 0.3 & 0.2 & 0.5\end{bmatrix} \tag{3-7}$$

The observation probability with respect to a state is

$$B=\left\{b_i\left(v_k\right)\right\}=\begin{bmatrix}0.3 & 0.7 & 0.3\\ 0.2 & 0.1 & 0.6\\ 0.5 & 0.2 & 0.1\end{bmatrix}=\begin{bmatrix}P_1(A) & P_2(A) & P_3(A)\\ P_1(B) & P_2(B) & P_3(B)\\ P_1(C) & P_2(C) & P_3(C)\end{bmatrix} \tag{3-8}$$

And the vector of initial state probability is

$$\pi=\begin{bmatrix}0.4 & 0.5 & 0.1\end{bmatrix}^T \tag{3-9}$$

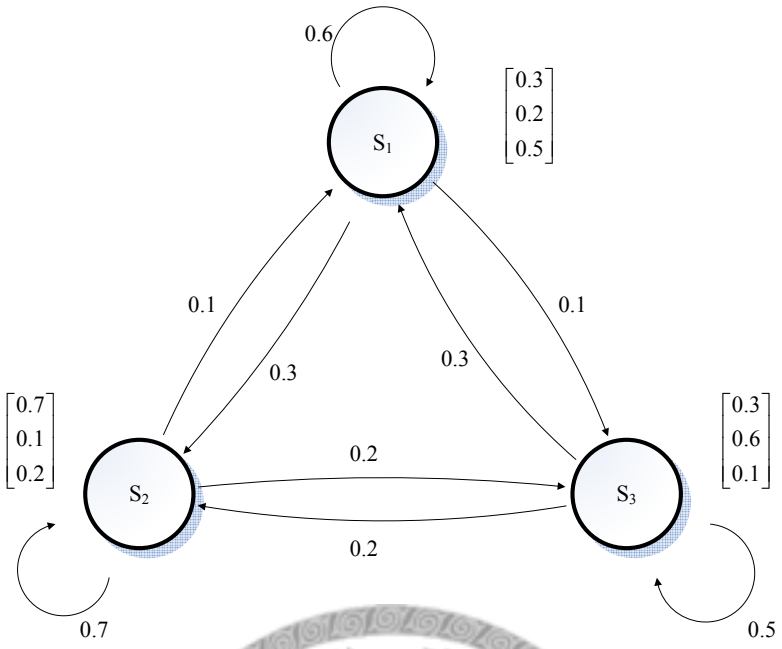

Fig. 3.2 Example of a three-state Hidden Markov Model.

Given an observation sequences $O=\{A,B,C,B\}$, there are 81 possible corresponding state sequences, and therefore the probability, $P(O|\lambda)$, is

$$P(O|\lambda)=\sum_{i=1}^{81}P(O,Q_i|\lambda)=\sum_{i=1}^{81}P(O|Q_i,\lambda)P(Q_i|\lambda),\ Q_i\text{: state sequence} \qquad (3\text{-}10)$$

For example, if states sequences $Q_i=\{S_2,S_2,S_3,S_1\}$, then

$$P(O|Q_i,\lambda)=P(A|S_2)P(B|S_2)P(C|S_3)P(B|S_1)=0.7*0.1*0.1*0.2=0.0014$$

$$P(Q_i|\lambda)=\pi(S_2)P(S_2|S_2)P(S_3|S_2)P(S_2|S_1)=0.5*0.7*0.2*0.3=0.021$$

The relative probability equals to $0.294*10^{-4}$. In fact it is not necessary to calculate all possible cases, it will be introduced in the below.

In the next step, there is a HMM model will be established. This process involves

computing probability of this model, training model by input data (six features that exact in above chapter), and finding the best sequence of state. We solve these problems in the next section.

3.2 Three Basic Problems for HMM

There are usually three basic problems that we have to solve using the HMM model. These problems are about evaluation efficiency, decoding, and training.

In the following sections, we describe several conventional solutions to these three standard problems.

3.2.1 The Evaluation Problem

The main concern in the evaluation problem is computational efficiency. Given a observation sequence $O = (o_1, o_2, ..., o_T)$, and a HMM model $\lambda = (A, B, \pi)$. The most straightforward way to compute $P(O \mid \lambda)$ is listing all possible state sequences and summing up their probabilities. It can be shown as:

$$P(O|\lambda) = \sum_{allQ} P(O, S|\lambda) = \sum_{allQ} \pi_{S_1} b_{S_1}(o_1) a_{S_1 S_2} b_{S_2}(o_2) ... a_{S_{T-1} S_T} b_{S_T}(o_T) \qquad (3\text{-}11)$$

Since the summation in (3-11) involves N^T possible Q sequences, the total

computational requirements are on the order of $2TN^T$ operations. The need to compute (3-11) without the exponential growth of computation, as a function of the sequence length T, is the first challenge for implementation of the HMM technique.

There is a more efficient method, called forward algorithm. First, define the forward variable:

$$\alpha_t(i) = P\left(O_1^t, q_t = S_i \middle| \lambda\right) \tag{3-12}$$

$\alpha_t(i)$ is the probability of the partial observation sequence $O_1^t = \{o_1, o_2, ..., o_t\}$ up to time t and state $q_t = S_i$ at time t. The forward variable can be calculated inductively by

$$\alpha_t(i) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\right] b_j(o_t) \qquad 2 \le t \le T, 1 \le j \le N \tag{3-13}$$

The desired result is simply

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3-14}$$

It requires about N^2T computations which are much less than direct calculation. Our application needs few N and more T (according to sentence length). This tremendous reduction in computation makes the HMM method attractive in our application.

3.2.2 The Estimation Problem

Given a set of observations $O=(o_1,o_2,...,o_T)$ and a HMM $\lambda=(A,B,\pi)$, the estimation problem involves finding the "right" model parameter values that specify a model most likely to produce the given sequence. This is often called “training” in speech processing.

In solving the estimation problem, we often follow the method of maximum likelihood (ML): Finding a model $\lambda=(A,B,\pi)$ such that $P(O|\lambda)$ is maximized for the given training sequence O. The problem can be solved by the iterative Baum-Welch algorithm. First we define the backward variable:

$$\beta_t(i)=P\left(O_{t+1}^T,q_t=S_i\middle|\lambda\right) \tag{3-15}$$

$\beta_t(i)$ is the probability of the partial observation sequence $O_{t+1}^T=(o_{t+1},o_{t+2},...,o_T)$, given state S_i at time t and the model λ. The backward

procedure can be set by:

(1) Initialization $\beta_T(i) = 1,\ 1 \le i \le N$ (3-16)

(2) Induction $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j),\ 1 \le t \le T\text{-}1,\ 1 \le j \le N$ (3-17)

Then, two new variables can be defined:

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) \quad (3\text{-}18)$$

$\xi_t(i,j)$ is the probability of being in state S_i at time t and state S_j at time $t+1$. And it can be inducted by forward variable and backward variable.

$$\xi_t(i,j) = \frac{P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)}{P(O \mid \lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{m=1}^{N}\sum_{n=1}^{N} \alpha_t(m) a_{mn} b_n(o_{t+1}) \beta_{t+1}(n)} \quad (3\text{-}19)$$

The variable $\gamma_t(i)$ is defined as

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) \quad (3\text{-}20)$$

Which is the probability of being in state S_i at time t. It also can be inducted by forward variable and backward variable.

$$\gamma_t(i) = \frac{P(O, q_t = S_i \mid \lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{3-21}$$

Based on an existing model λ, the Baum-Welch algorithm will be utilized to get the final models. After the variable are defined, the new parameters of HMM could be re-estimated as follows

$$\begin{aligned}\pi_i &= \text{expected number of times in state } S_i \text{ at time } (t=1) \\ &= \gamma_1(i)\end{aligned} \tag{3-22}$$

$$\begin{aligned}a_{ij} &= \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \\ &= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}\end{aligned} \tag{3-23}$$

$$b_j(v_k) = \frac{\text{expected number of times in state } S_j \text{ and observing symbol } v_k}{\text{expected number of times in state } S_j} \tag{3-24}$$

$$= \frac{\sum_{t=1,\ \text{s.t. } o_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

These parameters, then, would be updated, and go back to calculate the variables $\xi_t(i,j)$ and $\gamma_t(i)$. These operations will be repeated until parameters π, A, and B, are converged, and the HMM models are finally determinant.

The training process chooses the observation with single Gaussians having diagonal covariance. We trained the model using several minutes of speech data from reading textbook regularly with voicing states labeled in each frame.

3.2.3 The Decoding Problem

In this problem, the purpose is to find the best state sequence. The forward algorithm described in the previous section can not find out such a state sequence, and the Viterbi algorithm can be applied to solve this problem efficiently. The Viterbi algorithm can be regarded as the dynamic programming algorithm applied to the HMM or as a modified forward algorithm.

Instead of summing probabilities from different paths coming to the same

destination state, the Viterbi algorithm picks and remembers the best path. We define the probability quantity $\delta_t(i)$ which represents the maximum probability along the best probable state sequence path of a given observation sequence after t instants and being in state i.

$$\delta_t(i) = \max_{q_1, q_2, \cdots q_{t-1}} P\left(q_1 q_2 \ldots q_{t-1} q_t = S_i, O_t \mid \lambda\right) \tag{3-25}$$

And the best state sequence is backtracked by another function $\psi_t(i)$. Then the complete Viterbi algorithm can be described by using $\delta_t(i)$ and $\psi_t(i)$ by following steps:

Step 1: Initialization

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N \tag{3-26}$$

$$\psi_1(i) = 0 \tag{3-27}$$

Step 2: Induction

$$\delta_t(j) = \max_{1 \le i \le N}[\delta_{t-1}(i) a_{ij}] b_j(\mathrm{O}_t), \quad 2 \le t \le T,\ 1 \le j \le N \tag{3-28}$$

$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t\text{-}1}(i)a_{ij}], \quad 2 \le t \le T,\ 1 \le j \le N \tag{3-29}$$

Step 3: Termination

$$P^*(\mathbf{O}|\lambda) = \max_{1\le i\le N} \delta_T(i) \tag{3-30}$$

$$q_T^* = \arg\max_{1\le i\le N} \delta_T(i) \tag{3-31}$$

Step 4: Backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, ..., 1 \tag{3-32}$$

$$Q^* = (q_1^*, q_2^*, ..., q_T^*) \tag{3-33}$$

is the best sequence.

3.3 Performance

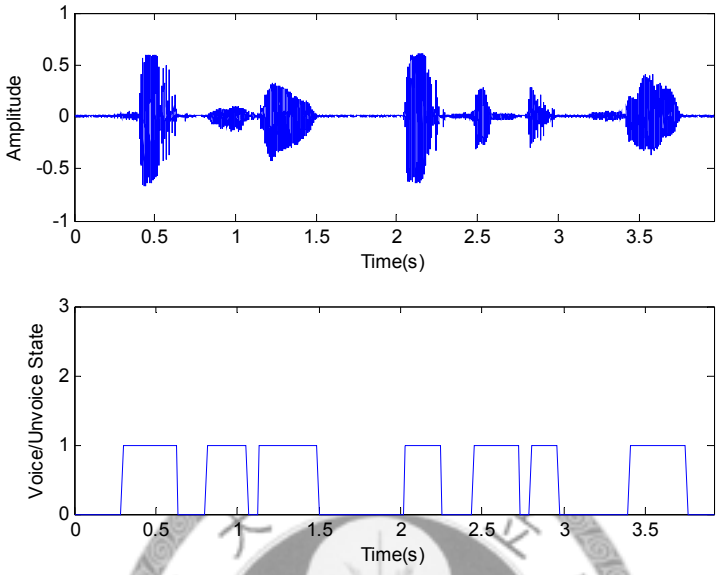

Fig.3.3 Performance of the model on speech

An example of the result that generalized by Viterbi algorithm for HMM is shown in Fig. 3.3.As we had hoped, the model has correctly indicated the word region by the voiced/uuvoiced states. The best states sequence are correspond to the frame index, defined as follow:

$$S_v(l) \cong Q(q_1, q_2, ..., q_T) \tag{3-34}$$

$S_v(l)$ is the state of the voice. This is an important feature that we will use in the

next chapter. At this point, we determine all of the nonverbal speech feature we were interested in. We can now use these feature on interaction between robot and people.

Chapter 4 Guided Oral Reading with Robot

Reading fluency is an indicator of reading ability. Improving reading ability can be approached by teaching the fluency of the sentence. Rhythm of a sentence is unique because of its syntax. Good comprehension of the syntax structure comes with the right rhythm in reading. In this chapter, the guided reading method with pet-like robot will be illustrated. Our target users will be the children, or the people who are lack of reading ability. We can get beginning and end time of each word after the HMM based endpoint detector from chapter 3. The robot dog can use this features to measure the reading fluency then giving an appropriate guidance to the user.

Before describing the fluency measurement method, the robot hardware will be introduced first, which included the illustration of robot appearance and function. The robot will be used to enhance the user reading ability.

Then the guided reading scenario can be designed for the user and robot. The tail mechanism plays a key role in this interaction. The detail of the tail swing mechanism will be showed in following section.

The turn-taking reading situation can be corresponded to the close loop control system. It was the basic concept of our guide reading process. To achieve this concept, the word phase model from the constant speed circular motion can be used in our

situation. The synchronization parameter and rhythm parameter can also be defined for describing the turn-taking reading condition. Then the final section illustrates the method for “controlling” the reader and improving the reading fluency.

4.1 Robotic Dog

The robotic dog is selected because its pet-like look is accessible for the children. The robot is about 90 cm high with four wheels and the friendly appearance (see the Fig. 4.1).

Fig. 4.1 The photo of the robotic dog

The hardware architecture is shown in Fig. 4.2.

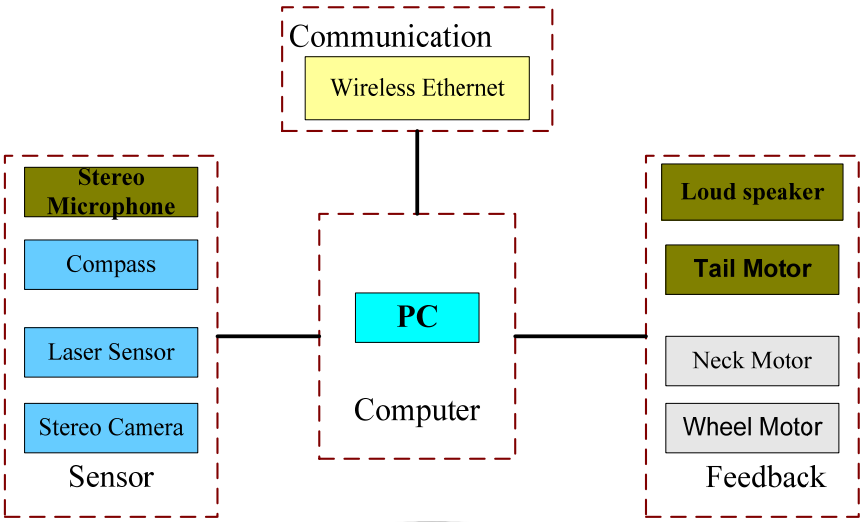

Fig. 4.2 The robotic dog hardware architecture

The control computer is the PC104 produced by Advantech. The laser sensor and the stereo camera are used to recognize the human face and measure the distance. And it uses the wireless Ethernet to communicate with the other devices. The microphone and speaker are devices for the guided reading application, with the help of the tail mechanism. Tail is acing as the mechanism for the guided reading interaction process, and it will be explained in detail in next section.

4.2 Interaction with Robot

The general way to teach children reading is the turn-taking reading, such that children repeating after the robot. With reading guide, turn taking process can be an effective and a powerful way to improve its reading fluency. Besides, our purpose is to

prepare an environment that children can learn by interacting with the robot, so we design an online training system with the robotic dog that can measure the reading fluency and giving the guidance while the user is reading.

Measuring reading speed, word accuracy, pause duration, pitch, and stress are necessary factors to assess oral reading fluency. We assume that user has the basic ability that he/she can recognize every single word of the article. He can read every single word slowly, but he may not actually know the phrase meaning. The word accuracy can be ignored by this assumption.

For simplifying the question, we also assume that the better pitch and stress of fluency can come with the better reading speed and the pause duration. Then the focus of this chapter is on the reading time adjustment.

On early childhood teaching in reading, the gesture or hand clapping are commonly used for guiding the correct word time for a child. Giving the word reading tempo can enhance the impression of the sentence rhythm. According to this idea, the tail mechanism is developed into a kind of metronome, or a kind of conductor's baton. Its speed and direction can be decided by controlling two motor, the photo of tail is shown in the Fig. 4.3.

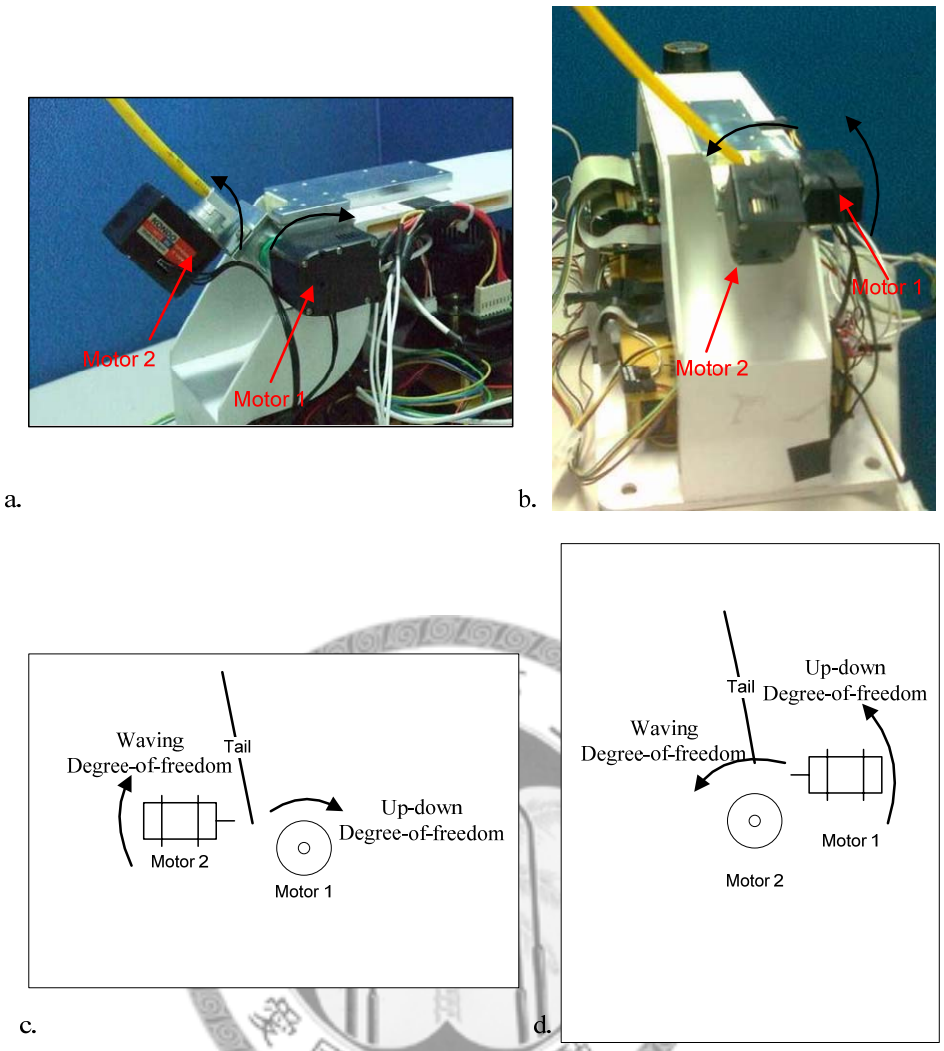

Fig. 4.3 The tail mechanism; (a), (b) show the photo of the tail mechanism. (c), (d) are diagrams of the tail corresponding to photos (a) and (b) respectively. The motor 1 control the degree-of-freedom of up and down. The motor 2 control the waving.

The tail can be used to point out the specific time of every word. Putting up and down can represent the beginning and the end of the guided reading. The waving time of the tail signals the each words time. The guided reading method with tail can be illustrated in Fig. 4.4.

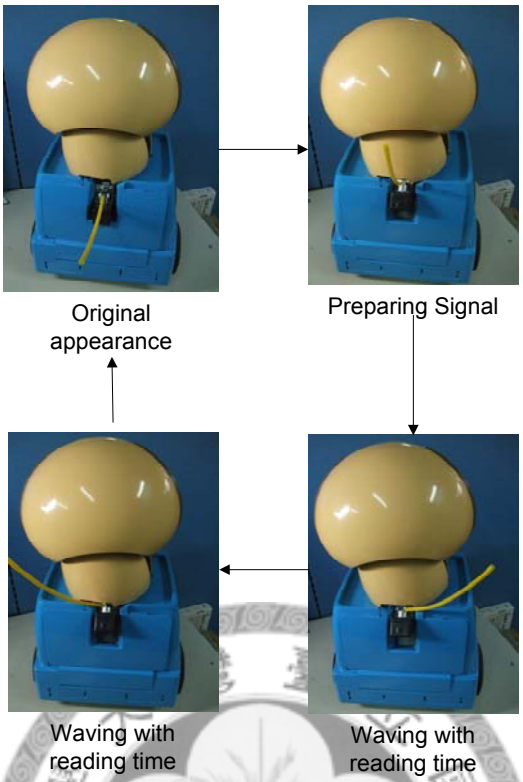

Fig. 4.4 Four conditions of the tail

Although it is impossible to catch the time exactly, the user generally can be lead to read in the specific rhythm by this way. The experiment in the next chapter verifies this point.

After a user reads the sentence, the voice signal can be analyzed and the the rhythm features could be calculated by the technique presented in the chapter 3. The process is under control to reach the desired fluency. The definition of the rhythm and the guiding signal provided b the tail will be illustrated in following section.

There is a scenario designed for teaching fluency with the robot. The flow chart of the guided reading is shown in Fig. 4.5. The user can be taught the right oral reading

paces of the article by the robot alone.

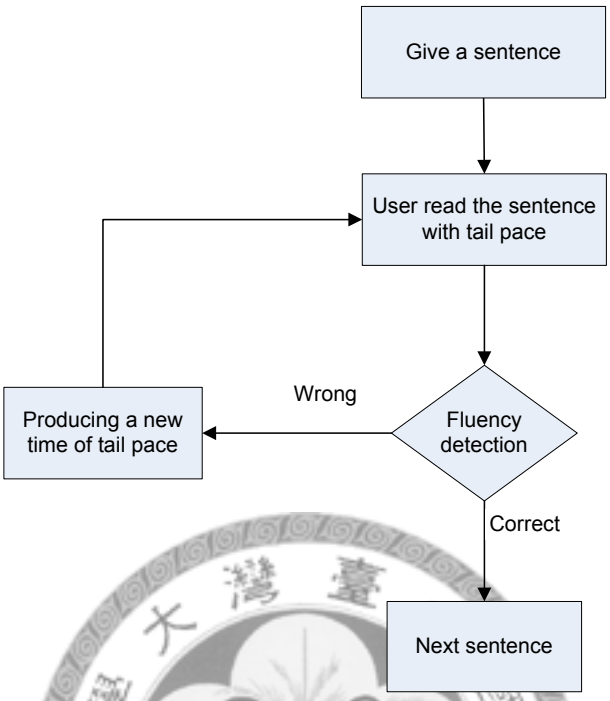

Fig. 4.5 Flowchart of the guided reading

4.3 Word Phase Control Model

In the guided reading process, the user can receive the signal which sent by the robot and response after the signal. Robot can give the new guidance signal according to the user output. This process is modeled by a closed-loop feedback system. A control system in general, can be represented by a reference input, controller, plant, and sensors. Controller produces input to the plant, or controlled system. The control error can be computed by comparing the reference and the measured output. Then the controller can generate a new system input. The closes-loop control system of the guided reading is

shown as the Fig. 4.6.

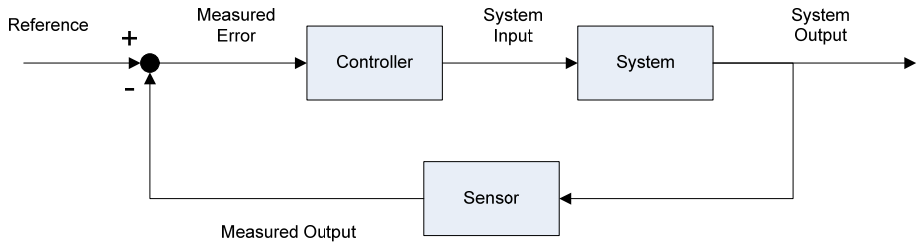

Fig. 4.6 Closed-loop control system

In this study, the user can be corresponded to the system, and the tail can be the controller. The robot is used to measure the error and produce the new waving time of the tail.

When the tail's signal is giving to user, the new output will be produced, just like the controller providing the signal to the system. Then the system output can be measured by the robot's sensor and assessed by the computer of robot. The new tail's action can be decided by this assessment of the fluency. The user can be affected by the different signals of the tail and read the sentence with different rhythm. So the Fig. 4.6 can be interpreted to the Fig. 4.7.

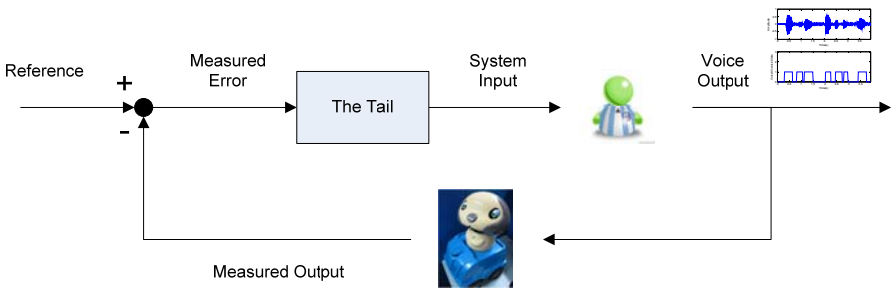

Fig. 4.7 Closed loop of the reading control system

To describe this idea, the output and the error have to be defined in control field. The mathematical description of the control process will be illustrated in the following section.

4.3.1 Phase Model

After the user reading a sentence, the robot can find out the boundary time of the every word, the reading speed and the pause duration can also be measured.

The mathematical description of the voice signal has to be concerned. Since the sentence is the basic unit of the guided reading process, all of the definition can be constrain in a single sentence.

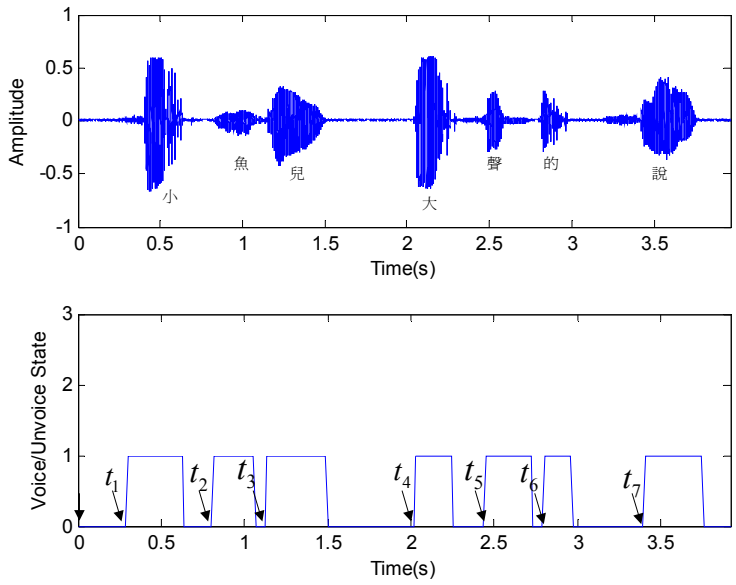

Fig. 4.8 Beginning time of words

The tail can wag with the defined time t_j. The t_j represents a teacher's rhythm

which is given in advance. It can lead the user to produce a voice signal. The feature can be extracted from voice signal as described in chapter 3. The time of the word beginning can be defined as t_k, which can be founded in the state of the voice, $S_v(l)$, as shown in Fig. 4.8. The variable N_w is defined as the number of the word of the sentence, $N_w = 7$ was used in the example.

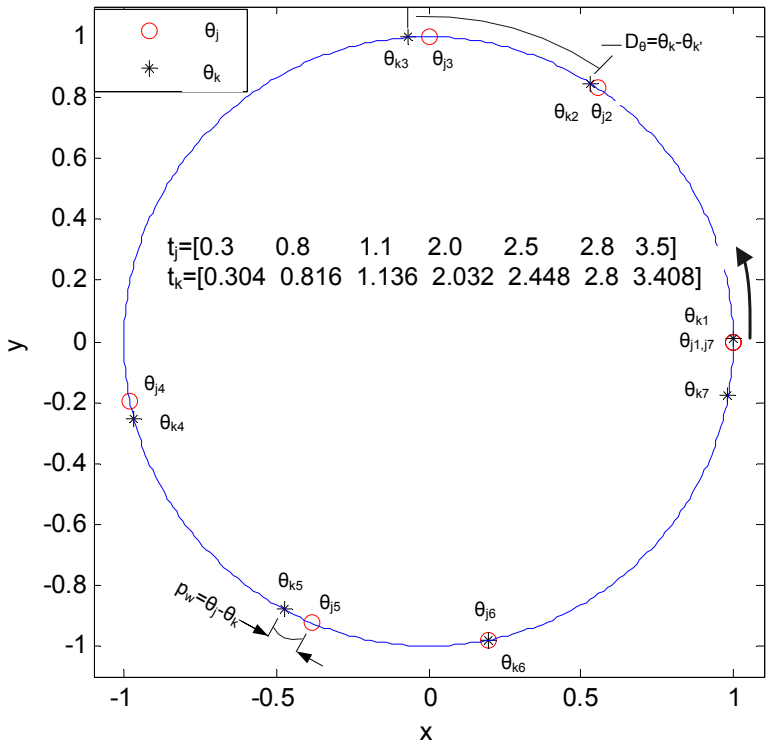

Fig. 4.9 Example for the phase model.

The control command time t_j and the user time t_k can be defined in a unit circle. θ_j and θ_k are phase angle corresponding to t_j and t_k, respectively. The p_w and D_θ are the synchronization error and the rhythm parameter.

The phase model was illustrated using an unit circle as shown in Fig. 4.9. The

standard time t_j can be treated as a particle which on the constant speed circular motion. The guided reading problem can then be transferred into the location control problem in terms of a unit circle's phase angle.

The t_j is the teacher's position on the unit circle. The position of the t_j can be defined as phase by the T_s, which means the teacher's sentence period.

$$T_s = t_j - t_1, \ when\ j = N_w \tag{4-1}$$

And the phase of the tail can be defined as

$$\theta_j = \frac{(t_j - t_1)}{T_s} * 2\pi \tag{4-2}$$

The output of the user's voice of a word can be treated as a location in the unit cycle that is to be controlled with guided motion commands. The word period can be defined in the circular domain too. The word phase can be defined as

$$\theta_k = \frac{(t_k - t_{j1})}{T_s} * 2\pi \tag{4-3}$$

The ω can be defined as the reading speed parameter.

$$\omega = \frac{2\pi}{T_s} \tag{4-4}$$

The tail waving “particle” and the user reading “particle” are also traveling along the circle with the ω. The controlled performance can be examined after finishing the each round of the guided reading by checking the location of the particle.

4.3.2 Error definition

In the guided reading process, the purpose is making sure about the user following the paces with the tail and reading with the rhythm of expectation. There have two steps of the error definition, synchronization parameter and the rhythm parameter.

Firstly, the synchronization error can be used to measure whether the user follow the tail paces exactly and the user is controllable with our method. The synchronization error can observe if the user follows the tail’s signal by calculating the difference between the θ_j and θ_k.

The synchronization error can be defined as

$$p_\theta = \sum_{j,k=1}^{N_k} \left| (\theta_k - \theta_j) \right| \tag{4-5}$$

The parameter p_θ means the error in the whole sentence, which can represent the

synchronization of the user and the tail paces.

The different users and different comprehension of the sentence may cause the different result. That will be an important parameter to determine the characteristic of the user.

It can't not be expected that the synchronization error be exactly zeros. There will be a specific error value to a specific user. The characteristic of a user in terms of the synchronization error must be decided by experiment.

The rhythm parameter can represent the preconception about the rhythm of the sentence. The rhythm error of the user is caused by his/her lack of understanding on the sentence.

Symbols for differentiating the different reading speed are illustrated here. For example:

<u>小 魚 兒</u> 大 聲 的。說

(1) Underline: Used in the phrase which we can read rapidly. That will be the fastest speed in the sentence.

(2) Space: Used in stressed phrase or the key word, which we can't read it slightly.

(3)The circle between two words: Use after the word we want to emphasis. The

pause will be longer than first two symbols.

To describe the reading speed, three levels of the pause duration are defined as follows. This rhythm variable is defined by the relative θ_k.

$$D = [\, D_1 \; D_2 \; D_3 \,] \tag{4-6}$$

where D_1, D_2, D_3 are defined as

$$D_1 = \frac{\sum \Delta\theta_k \;\; of \;\; the \; level \; 1}{number \;\; of \;\; the \; level \; 1} \tag{4-7}$$

$$D_2 = \frac{\sum \Delta\theta_k \;\; of \;\; the \; level \; 2}{number \;\; of \;\; the \; level \; 2} \tag{4-8}$$

$$D_3 = \frac{\sum \Delta\theta_k \;\; of \;\; the \; level \; 3}{number \;\; of \;\; the \; level \; 3} \tag{4-9}$$

It is anticipated that most sentences' pause duration could be described by the three-level pause duration parameters. When a sentence is given, the target D can be established by a fluent reader. Then the user's rhythm parameters are measured, and controlled to approach the target set up by the fluent reader.

4.3.3 Control Method

In case that the user can not make the goal due to poor following ability, the pace of the tail is modified according to the measured pause error. If the user's rhythm develops a bias to that of the tail, the guiding pace of the tail will be produced to eliminate the bias or error.

The command to the tail , $D^{'}$, can be calculated by the following equations.

$$D_1^{'} = \begin{cases} D_1 & , \ if \ D_{m1} \le D_1 \\ D_1 - (D_{m1} - D_1), & if \ D_{m1} > D_1 \end{cases} \quad (4\text{-}10)$$

$$D_2^{'} = \begin{cases} D_2 & , \ if \ D_{m2} \ge D_2 \\ D_2 + (D_2 - D_{m2}), & if \ D_{m2} < D_2 \end{cases} \quad (4\text{-}11)$$

$$D_3^{'} = \begin{cases} D_3 & , \ if \ D_{m3} \ge D_3 \\ D_3 + (D_3 - D_{m3}), & if \ D_{m3} < D_3 \end{cases} \quad (4\text{-}12)$$

Where the D_m is the measured rhythm parameters of user.

In case the user's pause duration was shorter than the tail's command, the control target has been reached. If the pause duration exceeds its targeted range, the control command will be re-calculated and adjusted for the next round in the sentence guiding

process.

Different user's characteristics may cause different controlling result. Consider p_θ as the controlled system characteristic, or the users' ability to follow the tail. Smaller p_θ indicates that the system has naturally smaller phase bias, which can also represent that the system can be controlled more easily, and the system is more stable to accept various ranges of control input. Considering the human ability to follow the tail pace, control equations can be further transformed into the following.

$$D_1^{'} = \begin{cases} D_1 & , \ if \ D_{m1} \le D_1 \\ D_1 - (D_{m1} - D_1) * \dfrac{(p_m - p_\theta)}{2\pi}, & if \ D_{m1} > D_1 \end{cases} \quad (4\text{-}13)$$

$$D_2^{'} = \begin{cases} D_2 & , \ if \ D_{m2} \ge D_2 \\ D_2 + (D_2 - D_{m2}) * \dfrac{(p_m - p_\theta)}{2\pi}, & if \ D_{m2} < D_2 \end{cases} \quad (4\text{-}14)$$

$$D_3^{'} = \begin{cases} D_3 & , \ if \ D_{m3} \ge D_3 \\ D_3 + (D_3 - D_{m3}) * \dfrac{(p_m - p_\theta)}{2\pi}, & if \ D_{m3} < D_3 \end{cases} \quad (4\text{-}15)$$

p_m represents the user's worst tracking error with respect to the tail paces, which can be measured by experiments. If the user has better ability to track tail paces, the more increment of amended command will be delivered with respect to the previous

command.

The following example explains the control procedure to improve a users reading rhythm. In this example, there is a sentence that the user wants to read it with better rhythm.

小 魚 兒 大 聲 的。說

The tail's default time may be t_j=[1.2 1.55 1.9 2.4 2.9 3.4 4.2] . The word phase can be calculated by Eq. 4-2, such that θ_j=[0 42 84 144 204 264 360] degrees. And the rhythm parameter can be calculated by Eq. 4-7 to 4-9. *D* was found to be *D*=[42 60 96] .

After the first round of the guided reading, we can get the user reading time, for example, we can get the time as: t_k =[1.21 1.61 2.03 2.51 3.02 3.5 4.35]. And the θ_k can also be calculated by Eq. 4-3. θ_k =[1.2 49.2 99.6 157.2 218.4 276 378] degrees. And the new D_m =[49.2 58.8 102]. Then p_θ can be calculate by Eq 4-7 to be p_θ = 81.6 . Then we assume the p_m is 180 for this user. So the D' can be calculated by Eq 4-13 to 4-15. The calculated results were shown as follows.

$$D_1^{'} = D_1 - (D_{m1} - D_1) * \frac{(p_m - p_\theta)}{360}$$

$$= 42-(49.2-42)*\frac{180-81.6}{360} = 40.032$$

$$D_2^{'} = 60+(60-58.8)*\frac{180-81.6}{360}=60.3280$$

$D_3^{'} = 96$, because $D_{m3} \geq D_3$

According to the $D^{'}$, the new waving time can be derived as,

$t_j^{'} = 1.2 \quad 1.5336 \quad 1.8672 \quad 2.3699 \quad 2.8727 \quad 3.3754 \quad 4.1754$

The new time has the difference between the original times, but not huge enough to confuse the user. The user was affected by this waving time and tended to get the right rhythm of reading. Different range of the command change must be adjusted and be accepted by different user. Magnitude of command changes can be tuned by setting the value of p_m.

Chapter 5 Experiment and Discussion

This chapter introduces some experiment results to illustrate the concept presented in Chapter 4. The purpose in the thesis was to design an autonomous robot which can be a company with children while they are reading by teaching the right rhythm of the sentence. Scenario of the experiment procedure was established as follows.

First of all, it has to be tested if the tail mechanism is effective to the user. The experiment can be done by simply testing the user's responses to various tail commands. The human response characteristic to the input device was obtained by step response tests. Finally, the reading fluency experiment with and without feedback control were compared, and the effectiveness of the proposed feedback method via robotic dog's tail mechanism was evaluated.

5.1 Experiment for Tail Mechanism Effect

The tail mechanism was the fundamental technique to "control" or guide the user response. In the first experiment, results from two situations were compared to show the tail effect. One of them is that the subject reads a randomly generated sentence with tail's guidance but only for the first and final word. Another reading experiment was carried out with the guidance for all of the words. And they also could watch the signal

at first, and read with the tail waving in the next time. Figure 5.1 shows the flow chart of these two experiments.

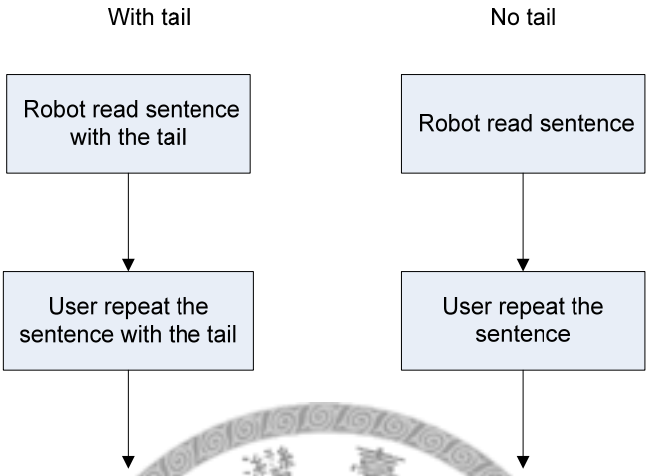

Fig. 5.1 The flow chart of experiment

The randomly generated sentence was used for ensuring that the word difficulty and syntax of sentence has minimum effects on the result. The error of the each word can be found by Eq. 5-1.

$$p_w = \frac{\theta_k - \theta_j}{2\pi}, \quad when\ k = j \tag{5-1}$$

The p_w of the 20 testing subjects was found and shown in the Figs. 5.2 and 5.3.

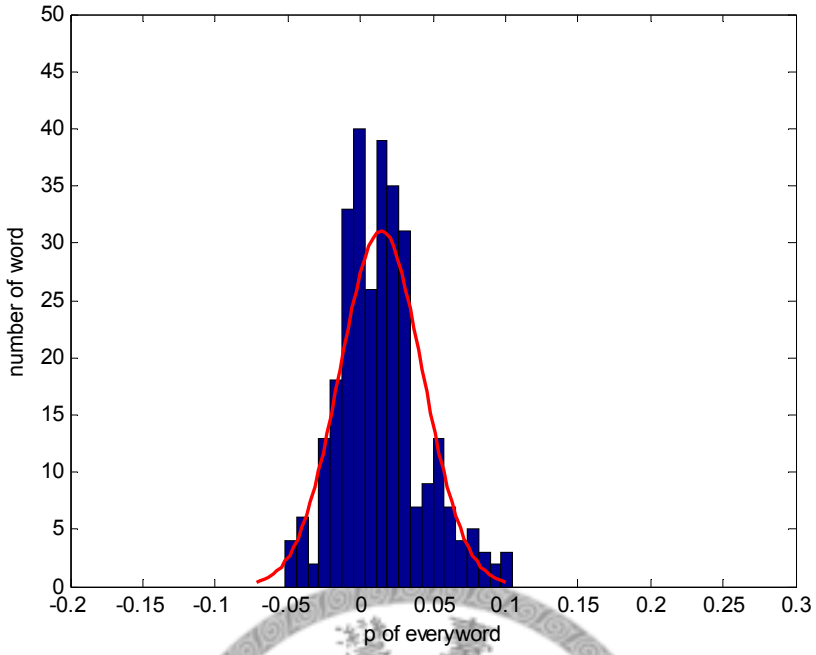

Fig. 5.2 p_w of the rhythm error with tail

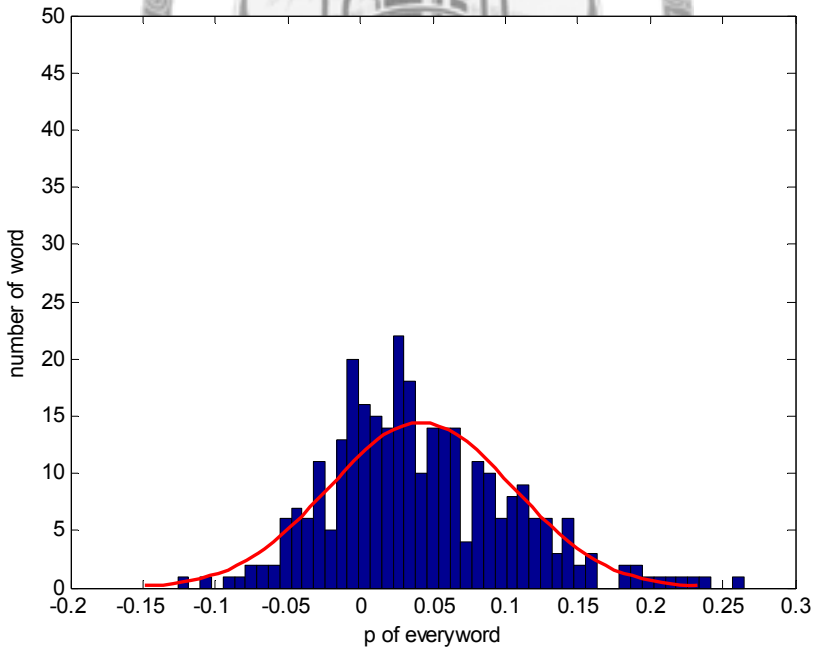

Fig. 5.3 p_w of the rhythm error without tail

The mean of the p_w is 0.0143 in Fig.5.1, and it is 0.0419 in Fig. 5.3. The standard deviations are 0.0287 and 0.0634, respectively. As we would expect, the tail is effective

in giving the rhythm command.

5.2 Experiment for Human Characteristic

In this section, human response to the input of the tail device was tested by a rapid step command of the tail. To obtain the step response of the user, instead of using turn-taking reading, the user has to read and learn the rhythm at the same time. The identification process measures the instinct reaction of the human user to the tail input, the response may indicate the inherent reaction speed of the test subject to the robotic dog's guiding device.

The process was repeated ten times with the same pace, and the p_θ was recorded as shown in the Fig. 5.4. And if the experiment process becomes that the waving time was changed every five reading rounds, the result has been recorded as shown in Fig. 5.5. Results shown in both figures were the averages of the subject's responses out of five experiments. The test sentences were randomly generated series of numbers.

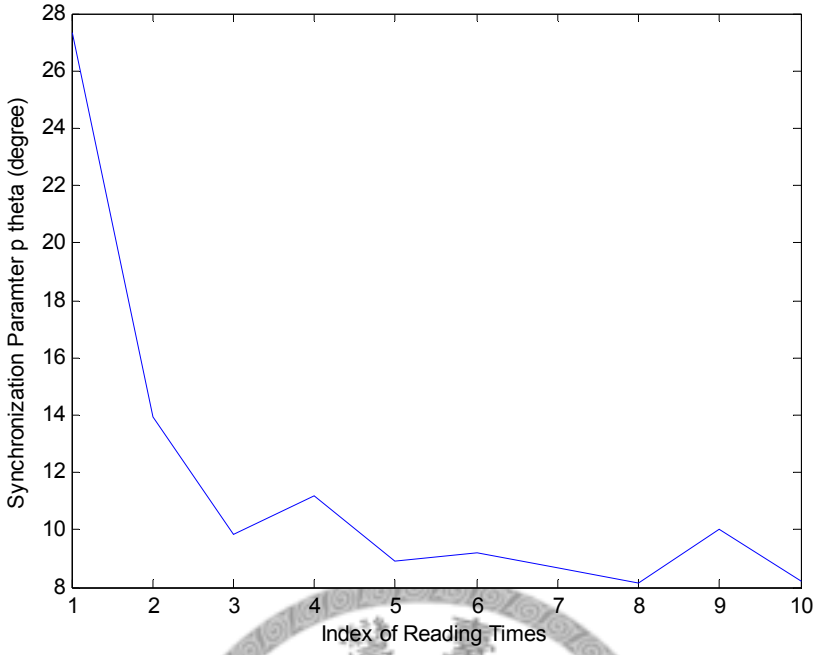

Fig. 5.4 Time evolution of the p_θ indicating the subject's ability to adapt to the input device.

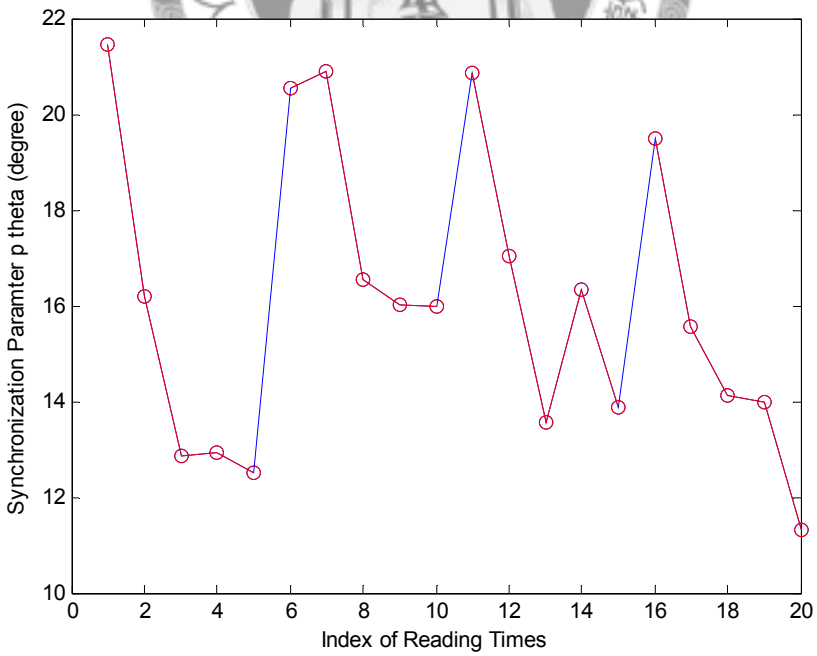

Fig. 5.5 p_θ performance showing the subject's response to step changes of input.

As we expected, the p_θ will drop straightly at beginning and maintain the

stability in a small range. p_θ can be used to identified subject's ability to adapt to a specific rhythm, and for identifying the subject's response characteristic. p_θ is different for different users, provides information about the worst possible human delay p_m, that is a parameter for controlling the tail pace.

5.3 Experiment of Turn-Taking Reading

The final experiment was expected to simulate the turn-taking reading situation. The robot demonstrated the ideal pace by the tail first, and the user follows the tail in the next time. In the next round, the tail's guiding time can be changed by the way Eqs. 4-13 to 4-15 described. The rhythm vector, D of one of the subject's test results were shown in Fig.5.6 and Fig. 5.7.

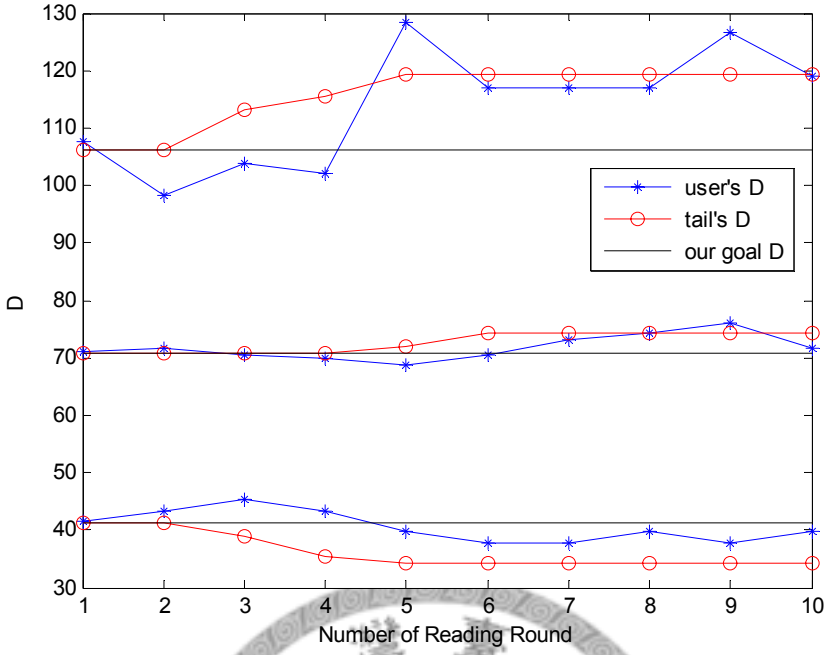

Fig. 5.6 Rhythm parameter response under control.

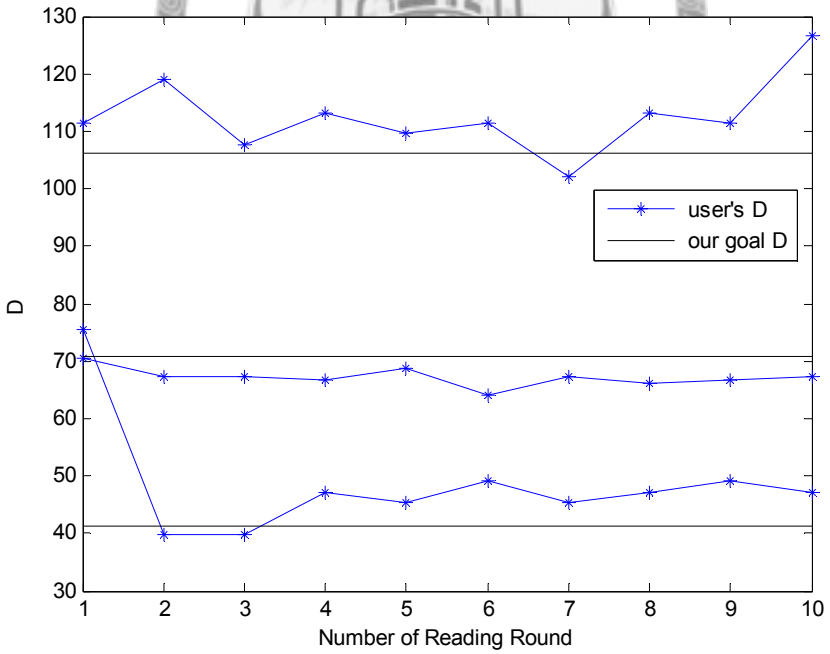

Fig. 5.7 Rhythm parameter response without control.

In this case, D=[41.31 70.82 106.23], and $p_m = 216$. It is shown that the rhythm of the oral reading con be effectively guided into the commanded values. As

guided time approached a large time, the results tended to be stable within a certain bound of the commanded rhythm.

Users with bigger p_m comes with results of the increasing overshoots and with more displacement from the original phase. This may cause the user's rhythm performance approaching the commanded rhythm values faster but also may cause the rhythm performance far from the original rhythm as shown in Fig. 5.8.

The figure 5.8 is two of the user's data of D_1 on the different p_m. It shows the effect of different p_m. Higher p_m results in bigger overshoots.

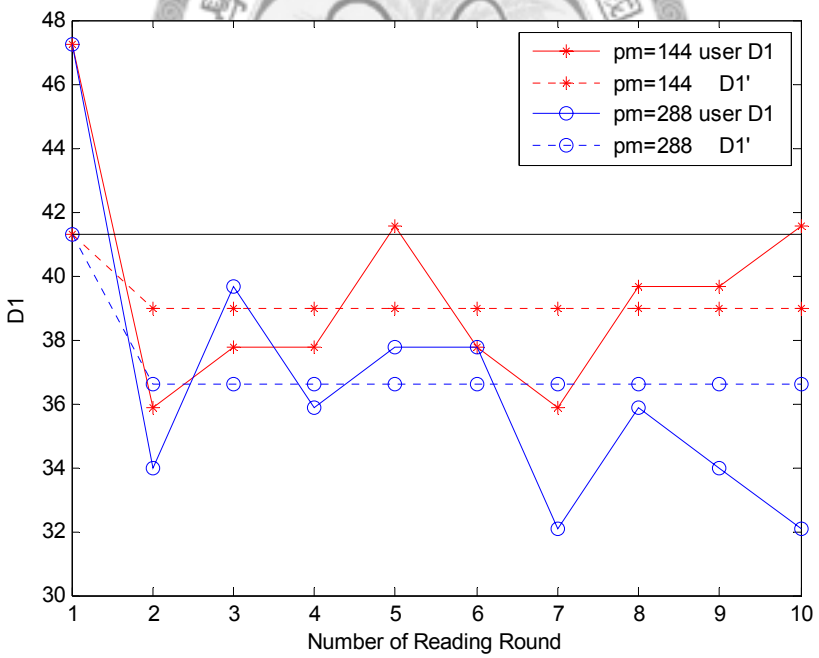

Fig. 5.8 The different p_m effect on the overshoots.

The ratio of the rhythm parameters indicate the user's ability to read fluently. Higher ratio of rhythm parameters expresses that the reader is able to manipulate the

sentence with high degree of freedom. In this respect, the ratio can be considered as the score of the oral reading performance. Here, two ratio parameters are defined as

$$C_2 = \frac{D_2}{D_1} \qquad (5\text{-}2)$$

$$C_3 = \frac{D_3}{D_1} \qquad (5\text{-}3)$$

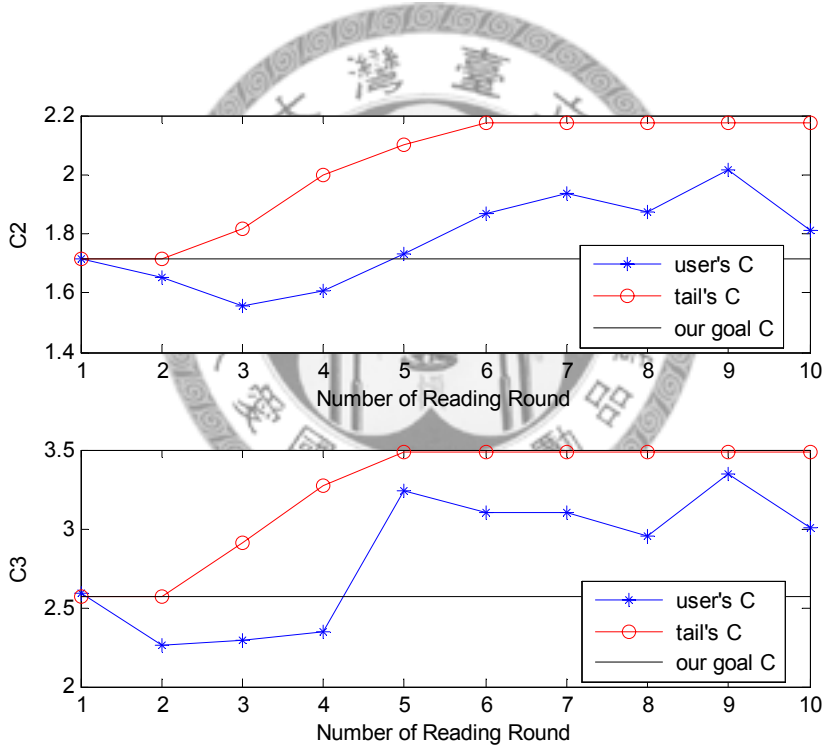

Fig. 5.9 Rhythm parameter ratio under control.

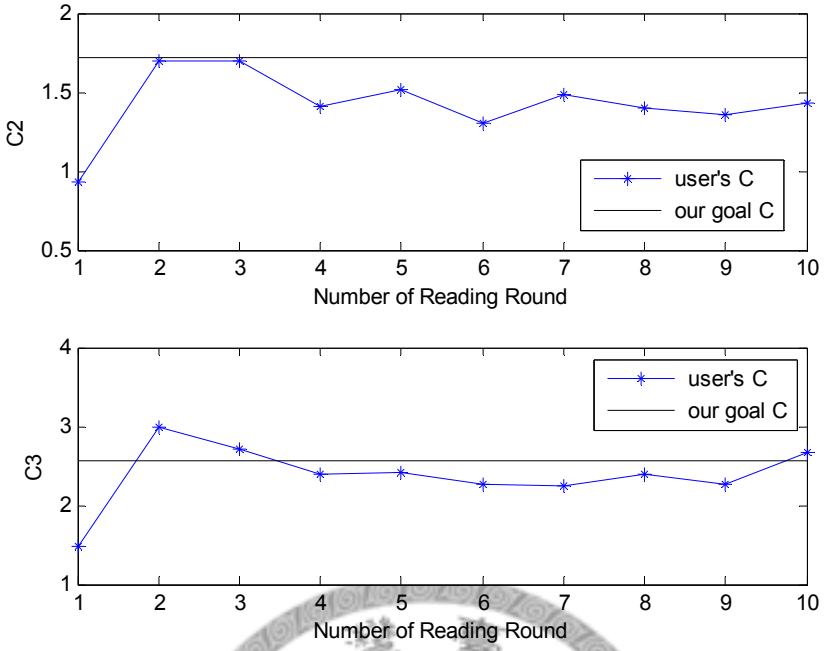

Fig. 5.10 Rhythm parameter ratio without control.

Figures 5.9 and 5.10 displayed the comparing results of the ratio of rhythm parameters trend. The user of the figure 5.9 reached the rhythm goal at the fifth times. And the user of the figure 5.10 doesn't reach the goal.

Data of ten subjects were recorded. The average time to reach the commanded ratio under control was 4.2th rounds, and the one without control was 9.2th round. The user's data with control reached the commanded ratios in a shorter time compared to readers not under control of the tail.

Chapter 6 Conclusions

Reading fluency is an indicator of reading ability. Improving reading ability can be approached by teaching the fluency of the sentence. Rhythm of a sentence is unique because of its syntax. Good comprehension of the syntax structure comes with the right rhythm in reading. In this study, the guided reading using a pet-like robot as an in-line feedback mechanism to readers was explored. The rhythmic behavior of the reader's oral reading voices were recognized and analyzed by the robot's computer. Beginning and ending time of each word was detected through a HMM based endpoint detector. Features of voices were then utilized to calculated error measures of the reading pace with respect to a teacher. Based on these error measures, rhythm parameters represent the reader's fluency while reading a sentence were displayed and control actions to improve the reading rhythm were given through the tail mechanism.

Main contribution of this study is the proposition that turn-taking reading guidance could be modeled by closed-loop control system. Each word in a single sentence is represented by its phase relationship with respect to the other words of the sentence. A constant speed circular motion was used to describe the sentence processing. A synchronization parameter and rhythm parameters were defined based on the error of phases between the phases of a teacher to that of a reader. Then the method for

‘controlling’ the readers reading rhythm and improving the reading fluency were provided.

The method was experimentally evaluated to illustrate its effectiveness. First of all, it has to be tested if the tail mechanism is effective to the user. The experiment can be done by simply testing the user’s responses to various tail commands. The human response characteristic to the input device was obtained by step inputs of rhythm commands. Reader’s controllability was modeled by a synchronization parameter. Based on this parameter, human’s delay in following the input commands was used to design control laws in the feedback loop to control the reading rhythm. The control law gives rhythm command changes following the error measures derived from the phase differences of each word in a sentence between a teacher and a reader’s oral reading voices. Results of reading fluency experiments with and without feedback control were compared, and the effectiveness of the proposed feedback method via robotic dog’s tail mechanism was evaluated.

It was shown that the proposed guided reading technique through a robot in the feedback control loop is feasible. Parameters defined in this study display clearly the reader’s oral reading fluency performance. Control strategies for improving the fluency were established with an error measure modeled by the phase difference of each reading word in a sentence conducted by a teacher and by the reader. It was shown by

experimental results that the reading rhythm is actually can be controlled and be improved while the reading is in progress. It was shown that reading fluency could be guided directly by the robot to reach in their rhythmic state in a few trials of a sentence.

Reference

[1] M. R. Kuhn, S. A. Stahl, " Fluency: a review of developmental Remedial practices," *Journal of Educational Psychology*, Vol. 95, No. 1, pp. 3-21, 2003.

[2] T. R. Rasinski, (2004)." Creating fluent readers," *Educational Leadership, 61,* 46–51

[3] M. A. Mastropieri, A. Leinart, & T. E. Scruggs, (1999). "Strategies to increase reading fluency," *Intervention in School and Clinic*, 34, 278–283.

[4] K. J. Topping, and G. A. Lindsay, (1992). "Paired reading: a review of the literature, " Research Papers in Education, 7, 3, 199-246.

[5] J.P. Brady, "Studies on the metronome effect on stuttering," Behav. Res. Ther. 7 (1969), pp. 197–204.

[6] N. Leonard, D. Paley, F. Lekien, R. Sepulchre, D. Fratantoni, and R. Davis, "Collective motion, sensor networks, and ocean sampling," *Proc. IEEE*, vol. 95, no. 1, pp. 48–74, Jan. 2007.

[7] 林葳葳, (1999), 名家教你朗讀, 國語日報出版社

[8] J.D. Freguson "Hidden Markov analysis: An introduction. In: Hidden Markov Models for Speech", (*Institute for Defense Analysis, Princeton*, 1980)

[9] S. Basu, "A linked-HMM model for robust voicing and speech detection," *Proc.*

ICASSP, 1, 816-819, 2003.

[10] J. Sohn, N.S Kim, W. Sung, "A statistical model-based voice activity detector," *IEEE Signal Process*. Lett. 6 (1) (January 1999) 1–3.

[11] M. Orlandi, A. Santarelli, and D. Falavigna. " Maximum likelihood endpoint detection with time-domain features." *In Proc of eurospeech*, pages 1757– 1760, 2003.