THE DEVELOPMENT OF A MUSIC READING AND SINGING TWO-WHEELED ROBOT

Wei-Chen Lee*, Hung-Yan Gu, Kuo-Liang Chung, Chyi-Yeu Lin, Chin-Shyurng Fahn,
Yah-Syun Lai, Chih-Cheng Chang, Chia-Lun Tsai, Kai-Jay Lu,
Huang-Liang Liau, and Mao-Kuo Hsu

ABSTRACT

The objective of this research was to construct a two-wheeled robot that can autonomously read music and sing songs with a synthesized voice. The music for the robot was created by using a musical notation editor developed in this study, which could be read by the robot's vision system. A musical notation recognition program converted the image obtained from the vision system to a text file, and then a voice synthesis program processed the text file to generate the synthesized music autonomously. We tested four Mandarin songs and the accuracy of recognition was over 98%. It demonstrated that a robot with integrated functions should be promising in entertainment applications in the future.

Key Words: music, reading, singing, voice synthesis.

I. INTRODUCTION

A singing voice usually can make people feel pleasant and comfortable, so it has become one of the indispensible elements in entertainment. A robot with singing capability may replace humans to serve the purpose. Someday when the techniques are mature, a robot may sing a solo in a theater to entertain its audience. Several singing robots have been proposed. Sony's QRIO (originally named SDR) has the capability of speech synthesis and singing voice production (Kuroki, 2006). It can also perform face recognition but cannot read music. Nakamura and Sawada built a mechanical voice system with auditory feedback learning capability to imitate a human vocalization (Nakamura and Sawada, 2006). Another singing robot, Pavarobotti, was created to sing songs by voice simulation (NCVS, 2008). The focus of previous research has been on voice synthesis only. Some other robots were able to perform singing. However, many of them sang songs by playing pre-recorded music.

The objective of this research was to build a mobile robot that can read music and sing songs by voice synthesis autonomously. Reading music printed on paper can increase the interaction between humans and robots, and singing songs by voice synthesis can enhance the sense of reality with greater flexibility. The proposed robot consists of a two-wheeled balancing platform, an image capturing system, a vision-based numbered musical notation recognition system, and a singing voice synthesis system as shown in Fig. 1. We will use an HP NX 6320 notebook computer with an Intel Core Duo 1.66 GHz CPU and 512 MB RAM to control these systems. Each system or program will be discussed in detail in the following sections.

II. TWO-WHEELED BALANCING PLATFORM

The finished robot is shown in Fig. 2. A two-wheeled balancing platform was constructed for the robot's base and body. The platform was composed of a controller, aluminum frames, a body case, two wheels, gear sets, timing belts, and lead-acid batteries.

*Corresponding author. (Tel: 886-2-27376478; Fax: 886-2-27376460; Email: wclee@mail.ntust.edu.tw)

W. C. Lee, C. Y. Lin and M. K. Hsu are with the Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan, R.O.C.

H. Y. Gu, K. L. Chung, C. S. Fahn, Y. S. Lai, C. C. Chang, C. L. Tsai, K. J. Lu and H. L. Liau are with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan, R.O.C.

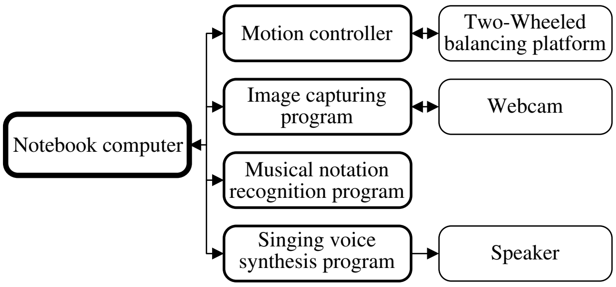

Notebook computer
Motion controller
Two-Wheeled balancing platform
Image capturing program
Webcam
Musical notation recognition program
Singing voice synthesis program
Speaker

Fig. 1 The schematic of the proposed robot

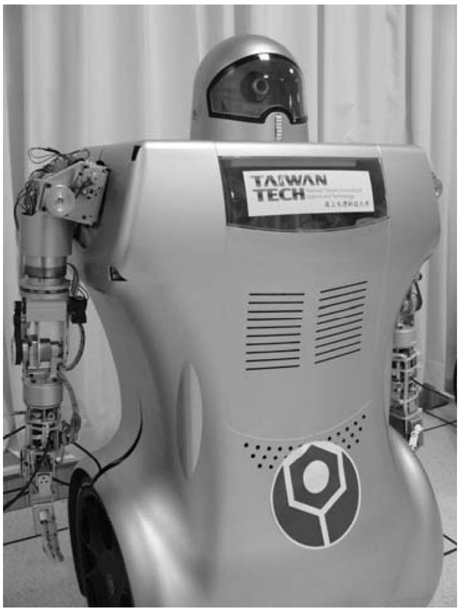


Fig. 2 The finished music reading and singing robot

The sensors used for keeping the platform in balance were a gyro sensor and a tilt sensor. The micro-controller embedded in the robot to control the movement and balancing was a PIC 18 microprocessor, which received the commands, such as go forward, balance, stop, etc., from the notebook computer and then performed the behavior according to the control law contained in it. The control law was obtained in the following way. First, the equations of motion of the robot were derived by using Kane's dynamics (Kim *et al.*, 2005). Then the equations were linearized about the equilibrium point and converted into the state-space form. Finally the PD (proportional-derivative) control was employed to complete the control law.

The experimental results for the robot's mobility tests showed that the robot can perform good balancing behavior, and can move forwards and backwards, make left turns and right turns, and stop. Fig. 3 shows the behavior of the robot when released around the neighborhood of the equilibrium point. It can balance itself very quickly and hold still with tiny swings.

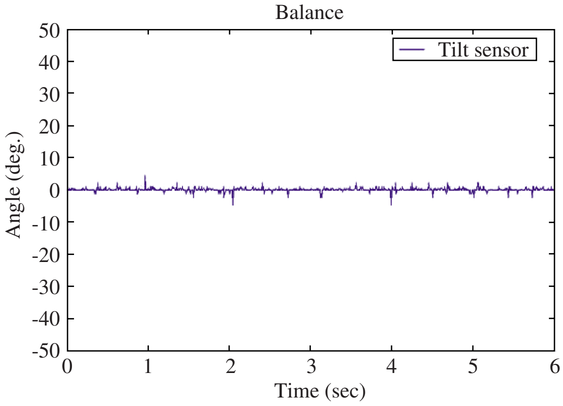


Fig. 3 The balancing of the two-wheeled robot

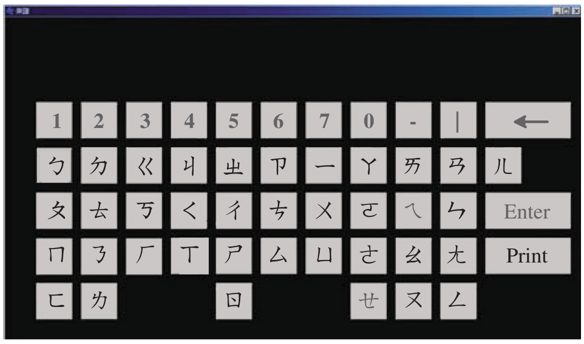


Fig. 4 The graphical user interface of the numbered musical notation editor

III. NUMBERED MUSICAL NOTATION EDITOR

To allow users to create or input a song easily, a musical notation editor was developed. The musical notation used was the numbered musical notation, which is popular in China, Taiwan and some Asian countries. The graphical user interface of the editor is shown in Fig. 4. To create the score of a song for robot use, a user needs to input the song by using the Mandarin phonetic symbols and the numbered musical notations. Then the editor automatically translates the input Mandarin phonetic symbols into the corresponding Pinyin and checks whether all the Chinese words and musical notations are processed. Finally the user can print out the score of the song created by the editor as shown in Fig. 5. The reason to use a score instead of a softcopy is to allow users to have more interactions with the robot and increase the sense of reality.

IV. IMAGE CAPTURING OF NUMBERED MUSICAL NOTATION

To capture an image of the musical notations as shown in Fig. 5, a webcam was installed in the head

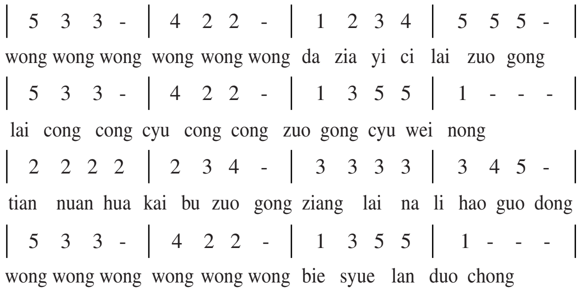

| 5 3 3 - | 4 2 2 - | 1 2 3 4 | 5 5 5 - |
wong wong wong wong wong wong da zia yi ci lai zuo gong
| 5 3 3 - | 4 2 2 - | 1 3 5 5 | 1 - - - |
lai cong cong cyu cong cong zuo gong cyu wei nong
| 2 2 2 2 | 2 3 4 - | 3 3 3 3 | 3 4 5 - |
tian nuan hua kai bu zuo gong ziang lai na li hao guo dong
| 5 3 3 - | 4 2 2 - | 1 3 5 5 | 1 - - - |
wong wong wong wong wong wong bie syue lan duo chong

Fig. 5 The output image

Fig. 6 The Logitech QuickCam Sphere MP webcam used in the robot

of the two-wheeled robot. The webcam we used was Logitech QuickCam Sphere MP of resolution 1280 × 960 as shown in Fig. 6, which can capture still images and video sequences. Note that the focus of the webcam in the robot system is fixed so manually adjusting the distance between the webcam and the music score may be needed.

Figure 7 presents the flowchart of the webcam-based image capturing system. At the beginning, the main program sends commands to the webcam to target on music notations, and then capture the image. After the webcam performs the image capturing task, a high resolution image of musical notations is generated to be recognized later. The major steps of the process are elaborated in the following.

1. Mark Location and Boundary Identification

For the convenience of finding a paper, a mark which could be easily recognized in a captured image was put on the upper-left corner of the paper. The mark we used here was our school's emblem. By means of color and geometry cues, this mark could be effectively identified from a complex background so as to help the webcam target on the paper rapidly.

After locating the mark, the main program started to identify the boundary of the paper from the neighborhood of the mark in a clockwise tracking sequence as shown in Fig. 8 by using the Sobel edge detector. In this study we attached the white paper with numbered music notations to black cardboard to make the boundary of the paper more obvious.

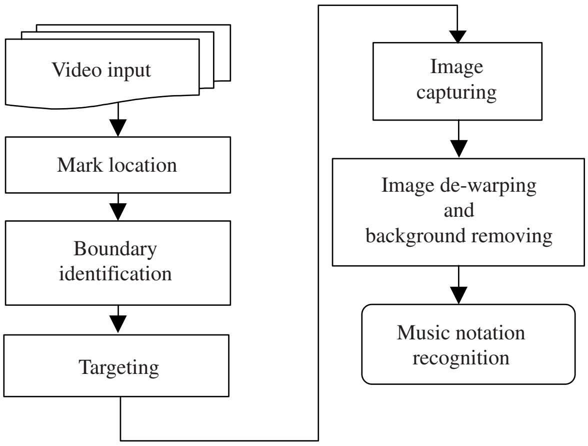


Fig. 7 Flowchart of the image capturing system

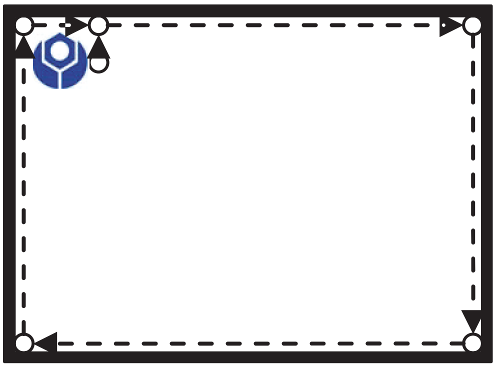


Fig. 8 The tracking sequence of boundary identification

2. Image De-Warping and Background Removing

Because the line of sight of the webcam is usually out of perpendicular to the paper with numbered musical notations, a novel de-warping method is needed for rectifying the skewed images captured by the webcam.

We restored the skewed rectangular image to a near rectangular one with the use of the method proposed by Li (1990). Then we applied an image interpolation method to it to increase the resolution. As a result, a high resolution image with numbered musical notations was produced.

3. Experimental Results

In the following experiment, a paper with numbered musical notations was put about 500 mm in front of the webcam. The first image captured by the webcam was skewed, with a complicated background, as shown in Fig. 9. After executing the entire capturing procedure described previously in this section, we obtained a high resolution image with numbered musical notations as demonstrated in Fig. 10.

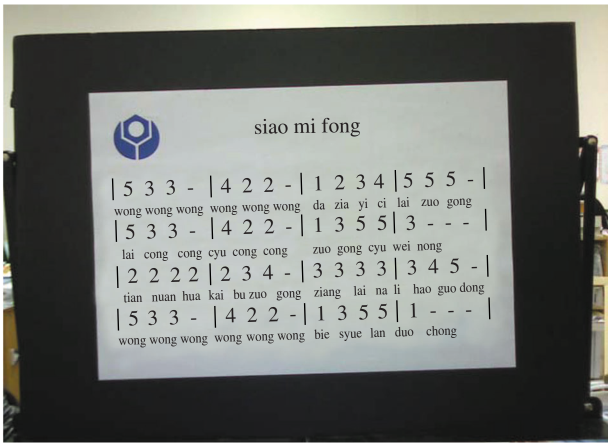

siao mi fong
|5 3 3 - |4 2 2 - |1 2 3 4|5 5 5 -|
wong wong wong wong wong wong da zia yi ci lai zuo gong
|5 3 3 - |4 2 2 -|1 3 5 5|3 - - -|
lai cong cong cyu cong cong zuo gong cyu wei nong
|2 2 2 2|2 3 4 -|3 3 3 3|3 4 5 -|
tian nuan hua kai bu zuo gong ziang lai na li hao guo dong
|5 3 3 - |4 2 2 -|1 3 5 5|1 - - -|
wong wong wong wong wong wong bie syue lan duo chong

Fig. 9 The image captured by the webcam

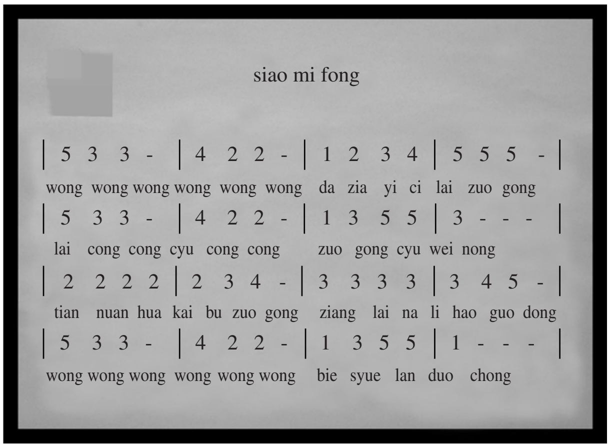
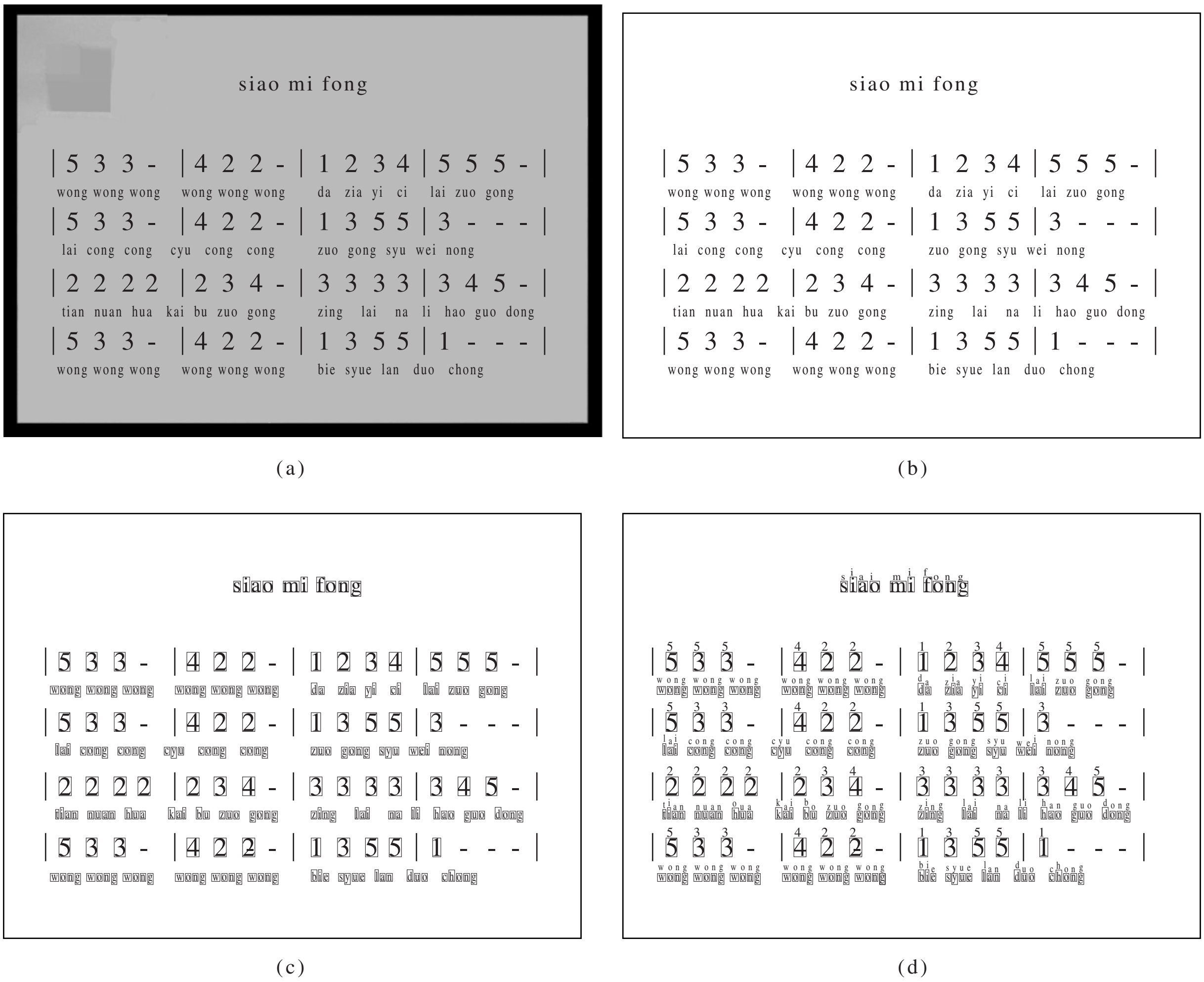

siao mi fong
|5 3 3 - |4 2 2 - |1 2 3 4|5 5 5 -|
wong wong wong wong wong wong da zia yi ci lai zuo gong
|5 3 3 - |4 2 2 -|1 3 5 5|3 - - -|
lai cong cong cyu cong cong zuo gong cyu wei nong
|2 2 2 2|2 3 4 -|3 3 3 3|3 4 5 -|
tian nuan hua kai bu zuo gong ziang lai na li hao guo dong
|5 3 3 - |4 2 2 -|1 3 5 5|1 - - -|
wong wong wong wong wong wong bie syue lan duo chong

Fig. 10 A high resolution image generated by the image capturing system


siao mi fong
|5 3 3 - |4 2 2 - |1 2 3 4|5 5 5 -|
wong wong wong wong wong wong da zia yi ci lai zuo gong
|5 3 3 - |4 2 2 -|1 3 5 5|3 - - -|
lai cong cong cyu cong cong zuo gong syu wei nong
|2 2 2 2|2 3 4 -|3 3 3 3|3 4 5 -|
tian nuan hua kai bu zuo gong zing lai na li hao guo dong
|5 3 3 - |4 2 2 -|1 3 5 5|1 - - -|
wong wong wong wong wong wong bie syue lan duo chong

(a)

(b)

(c)

(d)

Fig. 11 The images of each step in the proposed numbered musical notation recognition process: (a) The input rectified image; (b) the binary image; (c) the extracted blocks; (d) the recognized results

V. VISION-BASED NUMBERED MUSICAL NOTATION RECOGNITION

After obtaining the rectified image, we used a vision-based recognition system to recognize the numbered musical notations in the image. The main workflow of the proposed vision-based numbered musical notation recognition system is as follows. First, the Hough transform technique was applied to the rectified image as shown in Fig. 11(a) to locate the four corners (Chen and Chung, 2001). Second, the input gray-level image was converted to the binary image as shown in Fig. 11(b) by using Otsu's thresholding method (Otsu, 1979). Third, the binary image was partitioned into blocks by using a projection technique corresponding to the horizontal axis and vertical axis, respectively. Each block represents a symbol, which is either a numbered musical notation or a Latin alphabet character. The processed image is shown in Fig. 11(c). Finally, the corresponding numbered musical notations or Latin characters were recognized by comparing the content of each block with the model images stored in an established database. The recognized result is shown in Fig. 11(d). According to our tests, the proposed numbered musical notation recognition system was accurate. The accuracy will be discussed later in the paper. At the end, the recognized results were saved in a text file, which was used as the input of the singing voice synthesis system.

VI. SYNTHESIS OF MANDARIN SINGING VOICE

When a Mandarin song's score was generated by the vision recognition system, it was first parsed to extract each note's information. Then each note's lyric was used to load its corresponding HNM (harmonic-plus-noise model) parameters (Stylianou, 1996). HNM is based on a method of splitting the spectrum of a signal into two halves of unequal widths to better model the signal. The lower frequency half consists of harmonic partials while the higher frequency half consists of noise components. In terms of the HNM parameter values, a syllable can then be synthesized with higher clarity and naturalness level by using an extended HNM-based method proposed in this paper, which is an improvement on the original HNM method developed by Stylianou (1996; 2005). In the following we will be discussing some of the details of this proposed method.

1. How to Generate Mandarin Signals

The extended HNM method will mainly be used to process Mandarin songs. Mandarin is a syllable prominent language, and each syllable is of the structure, C_xVC_n (Chao, 1968). The initial part, C_x, may be null, a voiced consonant, or an unvoiced consonant while the final part, C_n, may be null or a nasal such as /*n*/. The kernel part, *V*, may be a vowel, diphthong, or triphthong. If the C_x was a short unvoiced consonant such as /*b*/, its synthetic signal was directly duplicated from the corresponding part in the recorded syllable. If the C_x was a long unvoiced consonant such as /*s*/, its synthetic signal was generated as a noise signal with HNM. Otherwise, the C_x was a voiced consonant (e.g., /*m*/) and was considered together with the remaining voiced phonemes, which were generated as harmonic partials plus noise signals with HNM.

```
HappySong  120  0.80
E3         1         cing
E3         1         tian
F3         1         gao
G3         1         gao
...
D3         1         cao
E3         1         er
F3         1         l
E3         1         wan
C3         1         yao
...
```

Fig. 12 The format of a score file

2. Score File Parsing and Note Data Interpretation

The song score stored in the text file generated by the vision system is shown in Fig. 12. The first line was song's name, tempo (e.g., 120 means 120 beats per minute), and fullness ratio (e.g., 0.80 means only 80% of a note's duration is used for singing and the rest 20% is reserved for transiting to the following note). Each of the remaining lines contained a note's pitch symbol, number of beats, and lyric.

After all notes' pitch frequencies were determined, automatic key shifting was executed. Key shifting must be done to allow the pitch range of the score file to match the pitch range of the person who uttered the source Mandarin syllables beforehand for analyzing HNM parameters. Otherwise, if the music notes are very high in pitch, we may synthesize very high pitch voices that regular people cannot generate. In this research, key shifting was done in the following steps: (a) Find the maximum and minimum values from the notes' pitch frequencies; (b) Calculate the average of the maximum and minimum values; (c) Compute the ratio of the person's analyzed mean pitch frequency to the average value; (d) Multiply each note's pitch frequency with the ratio.

3. Synthesis of Signal Waveform

Two issues have not been explicitly solved in the literature on HNM. The first issue is how to warp the time axis of a synthetic syllable so that more fluent syllable signals can be synthesized. When a syllable's duration needs to be changed, a simple time warping method, i.e., linear warping, will usually result in lower fluency. The second issue is how to keep the timbre of synthetic syllables consistent. Timbre is the characteristics in a voice, which allows people to identify someone's voice from others'. The first issue will be addressed in subsection 1 and the second issue in subsections 2 and 3 below.

(i) Planning of Phoneme Duration

For a short unvoiced phoneme such as /*b*/, its time length was planned as the corresponding phoneme length in the recorded syllable. However, for a long unvoiced phoneme such as /*s*/, its length was planned by multiplying its original length with a factor, which is usually confined to within the range from 0.6 to 1.4.

Regarding the voiced phonemes of a syllable, e.g. /*m*/, /*a*/, and /*n*/ in /man/, the phoneme durations were planned differently. It was observed that the consonant-to-vowel duration ratio becomes smaller when a syllable is sung. By decreasing the lengths of the leading and following consonants, we could have a long enough duration for the kernel vowel. After the values of the durations were determined, a mapping function from the phonemes in the synthetic syllable to the corresponding phonemes in the recorded syllable could then be established. An example of a piecewise linear time mapping function for the syllable /man/ is shown in Fig. 13.

(ii) Determination of Pitch-Tuned HNM Parameters

According to the constructed mapping function, an analyzed frame's time position on a recorded syllable's time axis could then be mapped to a time position, also called a control point, on a synthetic syllable's time axis. On the control point, the pitch-original (carrying the pitch of the person who uttered the syllable) HNM parameters, A_i(amplitude), F_i (frequency), and θ_i(phase), for the i-th harmonic partial could be obtained by referring to its corresponding analysis frame. However, the parameters, $\tilde{A}_k$, $\tilde{F}_k$, and $\tilde{\theta}_k$, for a pitch-tuned (to have the pitch of the corresponding note) harmonic partial should be determined carefully to keep the synthetic voice timbre consistent (i.e., timbre was not changed as the pitch shifted).

A principle to achieve this is to keep the spectral envelope unchanged (Dodge and Jerse, 1997). This implies that the amplitude $\tilde{A}_k$ of the pitch-tuned harmonic partial located at frequency $\tilde{F}_k$ must be interpolated according to the spectral envelope defined by the sequence of pairs, (F_i, A_i). In detail, for the k-th harmonic frequency $\tilde{F}_k$, we first found a pitch-original harmonic frequency F_j that was nearest to but less than $\tilde{F}_k$. Then, the four pitch-original harmonic partials of the frequencies, F_{j-1}, F_j, F_{j+1}, and F_{j+2}, were used to perform order-three Lagrange interpolation (Faires and Burden, 1998) according to Eq. (1) to compute the value of $\tilde{A}_k$.

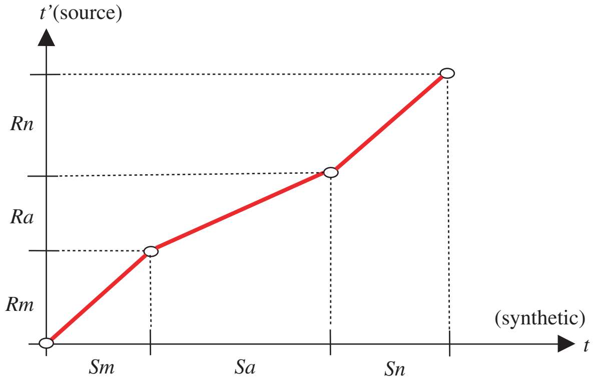


Fig. 13 An example of time mapping function

$$\tilde{A}_k = \sum_{m=j-1}^{j+2} A_m \times \prod_{\substack{h=j-1 \\ h\neq m}}^{j+2} \frac{\tilde{F}_k - F_h}{F_m - F_h}\,. \tag{1}$$

The phase $\tilde{\theta}_k$ was interpolated in the same way.

(iii) Synthesis of Signal Samples

For the harmonic signal, $H(t)$, between the n-th and $(n + 1)$-th control points, its sample values are computed as

$$H(t) = \sum_{k=0}^{L} a_k^n(t)\cos(\phi_k^n(t))\,, \quad t = 0, 1, \cdots, T^n, \tag{2}$$

$$a_k^n(t) = \tilde{A}_k^n + \frac{t}{T^n}(\tilde{A}_k^{n+1} - \tilde{A}_k^n)\,, \tag{3}$$

$$\phi_k^n(t) = \phi_k^n(t-1) + 2\pi f_k^n(t)/22050,\ \phi_k^n(0) = \hat{\theta}_k^n, \tag{4}$$

$$f_k^n(t) = \tilde{F}_k^n + \frac{t}{T^n}(\tilde{F}_k^{n+1} - \tilde{F}_k^n)\,, \tag{5}$$

where L is the number of harmonic partials, T^n is the number of samples between the n-th and $(n + 1)$-th control points, 22050 is the sampling rate we chose, $a_k^n(t)$ is the time-varying amplitude of the k-th partial at time t from the start of the n-th control point, $\phi_k^n(t)$ is the cumulated phase for the k-th partial, $f_k^n(t)$ is the time-varying frequency for the k-th partial, and $\hat{\theta}_k^n = puw(\tilde{\theta}_k^n, \hat{\theta}_k^{n-1})$, i.e., unwrapped phase of $\tilde{\theta}_k^n$ versus $\hat{\theta}_k^{n-1}$.

We synthesized noise signals as summations of sinusoidal components. Let G_k be the frequency of the k-th sinusoid. Its value is set as $G_k = 100 \times k$ (Hz) (Stylianou, 1996). Let B_n^k be the noise amplitude for the k-th sinusoid on the n-th control point. To determine its value, the 10 cepstrum coefficients of the n-th control point representing the noise spectral envelope were first appended with zero values and inversely transformed to the spectral domain by using inverse discrete Fourier transform. Then, exponentiation was taken to obtain the corresponding spectral magnitude coefficients, X_j, $j = 0, 1, \cdots, 2047$. According to the magnitudes X_j, the value of B_n^k was

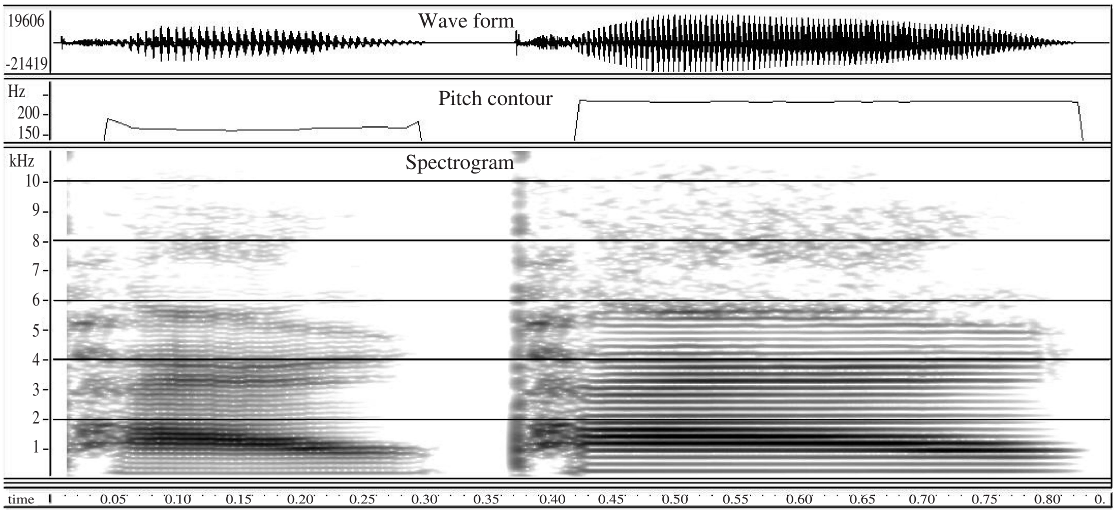


Fig. 14 Waveforms, pitch contours, and spectrograms for the syllable /pao/

obtained by linearly interpolating the two adjacent magnitudes, X_i and X_{i+1}, whose frequencies surround the frequency of G_k.

An example of the synthetic syllable /pao/ is shown on the right side of Fig. 14 while the originally recorded /pao/ is on the left side. The top chart is the signal waveform, the middle chart is the pitch contour, and the bottom chart is the spectrogram. Fig. 14 shows that the synthetic syllable was lengthened while its pitch was elevated as compared to the recorded one. Also, in the spectrogram of the synthetic syllable, the horizontal parallel lines represent harmonic partials that were synthesized with Eqs. (2), (3), (4), and (5). As to the cloud-shaped strips, they represented noises and were synthesized as summations of sinusoidal components.

The proposed method can synthesize a Mandarin singing voice efficiently. It took 0.32 second of CPU time on average to synthesize a syllable of 1.0 second in length. The synthetic signal is very clear and fluent.

VII. EXPERIMENTAL RESULTS AND DISCUSSION

All the subsystems of the robot were connected together through the notebook computer, and the robot was ready to perform. The performance we proposed was as follows. First, the song was prepared by using the editor developed in this study. Then the robot was controlled to stop in front of the music stand where the song score was put as shown in Fig. 15. Then the robot autonomously captured the image, recognized the numbered musical notations and then sang the song with synthesized voice.

The recognition accuracy was used as a measure of the performance of the numbered musical notation recognition system in the robot. The calculation of the recognition accuracy is given by $(t\text{-}w)/t \times 100\%$, where t denotes the total number of the symbols, which include Latin characters and numbered musical notations; w denotes the average number of incorrect symbols after executing our recognition process 10 times. Based on the four Mandarin songs we tested, the recognition accuracy of the proposed numbered musical notation recognition system was obtained. Figs. 16(a)-(d) show the images of the four Mandarin songs, Siao Mi Fong, Siao Sing Sing, Wo Ai Tai Mei, and Cian Li Zhii Wai, respectively. Table 1 shows that the average recognition accuracy of the proposed numbered musical notation recognition system was over 98%, which demonstrates that the proposed numbered musical notation recognition system can achieve high recognition accuracy.

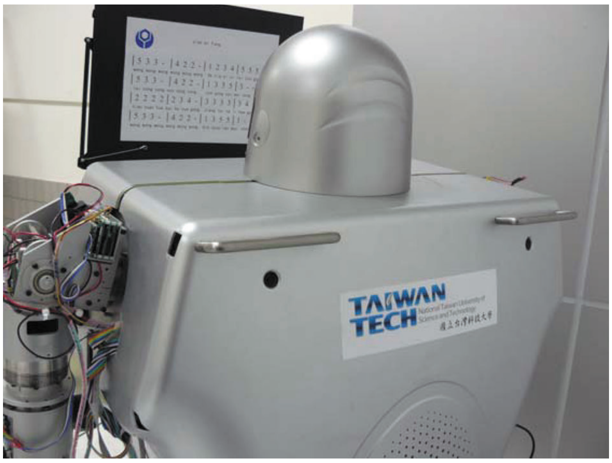


Fig. 15 The robot is reading the musical notations of the song and then singing it

Table 1 The average recognition accuracy of the recognition system

	Number of characters	Average incorrect recognized characters per run	Average recognition accuracy (%)
Siao Mi Fong	363	0.1	99.97
Siao Sing Sing	230	3.9	98.30
Wo Ai Tai Mei	171	0.2	99.88
Cian Li Zhii Wai	139	0.7	99.50

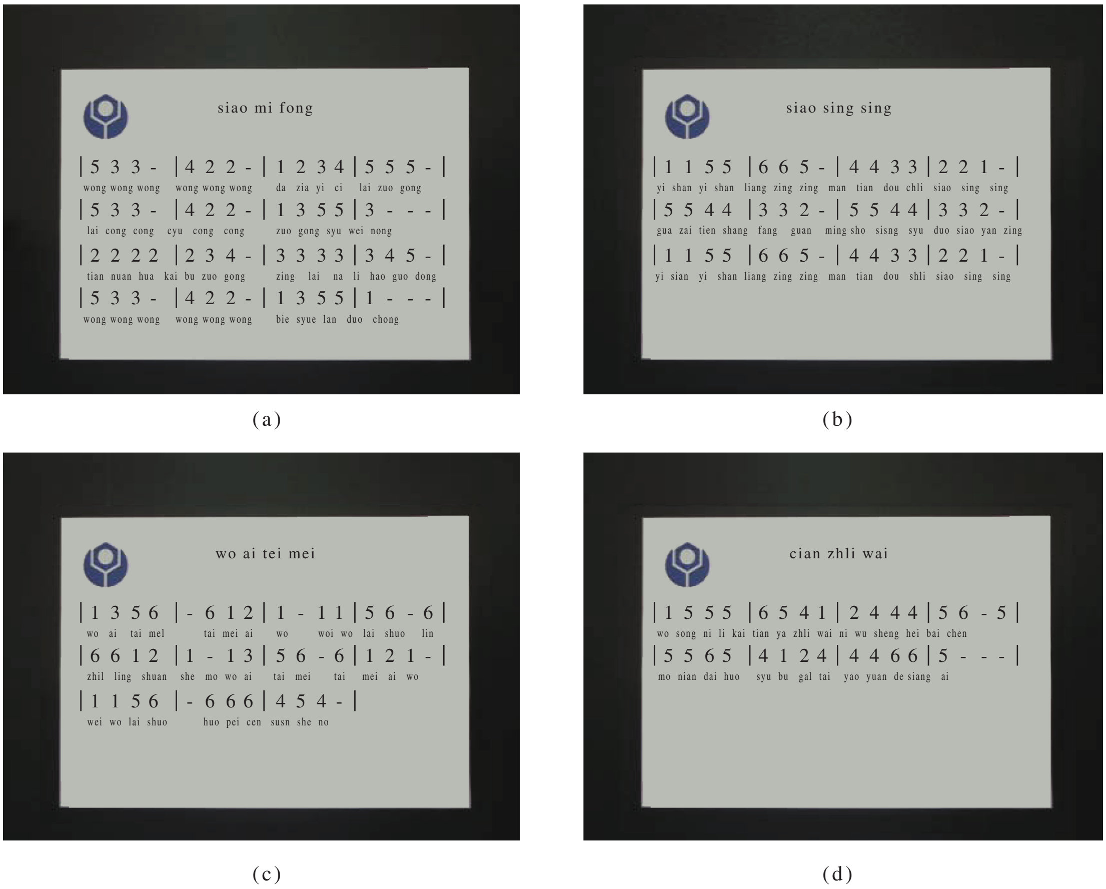

siao mi fong
| 5 3 3 - | 4 2 2 - | 1 2 3 4 | 5 5 5 - |
wong wong wong wong wong wong da zia yi ci lai zuo gong
| 5 3 3 - | 4 2 2 - | 1 3 5 5 | 3 - - - |
lai cong cong cyu cong cong zuo gong syu wei nong
| 2 2 2 2 | 2 3 4 - | 3 3 3 3 | 3 4 5 - |
tian nuan hua kai bu zuo gong zing lai na li hao guo dong
| 5 3 3 - | 4 2 2 - | 1 3 5 5 | 1 - - - |
wong wong wong wong wong wong bie syue lan duo chong
(a)
siao sing sing
| 1 1 5 5 | 6 6 5 - | 4 4 3 3 | 2 2 1 - |
yi shan yi shan liang zing zing man tian dou chli siao sing sing
| 5 5 4 4 | 3 3 2 - | 5 5 4 4 | 3 3 2 - |
gua zai tien shang fang guan ming sho sisng syu duo siao yan zing
| 1 1 5 5 | 6 6 5 - | 4 4 3 3 | 2 2 1 - |
yi sian yi shan liang zing zing man tian dou shli siao sing sing
(b)
wo ai tei mei
| 1 3 5 6 | - 6 1 2 | 1 - 1 1 | 5 6 - 6 |
wo ai tai mel tai mei ai wo woi wo lai shuo lin
| 6 6 1 2 | 1 - 1 3 | 5 6 - 6 | 1 2 1 - |
zhil ling shuan she mo wo ai tai mei tai mei ai wo
| 1 1 5 6 | - 6 6 6 | 4 5 4 - |
wei wo lai shuo huo pei cen susn she no
(c)
cian zhli wai
| 1 5 5 5 | 6 5 4 1 | 2 4 4 4 | 5 6 - 5 |
wo song ni li kai tian ya zhli wai ni wu sheng hei bai chen
| 5 5 6 5 | 4 1 2 4 | 4 4 6 6 | 5 - - - |
mo nian dai huo syu bu gal tai yao yuan de siang ai
(d)

Fig. 16 The images of the four Mandarin songs: (a) Siao Mi Fong; (b) Siao Sing Sing; (c) Wo Ai Tai Mei; (d) Cian Li Zhii Wai

The recognition accuracy was apparently affected by the lighting conditions. In a well-lighted environment, the recognition accuracy can be further increased. Regarding the voice synthesis, the synthetic voice sounded very clear and fluent, but still slightly different from the voice generated by a real human. More work will be needed to improve this situation.

VIII. CONCLUSIONS

A two-wheeled balancing robot that can read music and sing songs autonomously using voice synthesis was constructed in this study. In contrast to most of the developed robots which have one or two major functions, the robot research focused on the integration of all the subsystems, including a self-balancing platform, a music notation editor, an image capturing system, a numbered music notation recognition system, and a voice synthesis program. People can create a song by using the music notation editor developed in the study and show it to the robot. Then the robot can make a performance by using its vision, music notation recognition, and voice synthesis capabilities. The HNM technique was improved in this study to synthesize the voice. Experimental results showed that on average, the recognition accuracy for the robot is over 98%, which demonstrated that to have such a robot, with integrated functions, perform in a theater to entertain people, will probably be feasible in the near future.

ACKNOWLEDGMENTS

The authors wish to thank the National Science Council of the Republic of China for financial support of the research under grant NSC-95-2218-E-011-009.

NOMENCLATURE

A_i amplitude for the *i*-th pitch-original harmonic partial

$\tilde{A}_k$ amplitude for the *i*-th pitch-tuned harmonic partial

$a_k^n(t)$ time-varying amplitude of the *k*-th partial at time *t* from the start of the *n*-th control point

B_n^k noise amplitude for the *k*-th sinusoid on the *n*-th control point

F_i frequency for the *i*-th pitch-original harmonic partial

$\tilde{F}_k$ frequency for the *i*-th pitch-tuned harmonic partial

$f_k^n(t)$ time-varying frequency for the *k*-th partial

G_k frequency of the *k*-th sinusoid

$H(t)$ harmonic signal

L number of harmonic partials

T^n number of samples between the *n*-th and (n + 1)-th control points

t time, in second

θ_i phase for the *i*-th pitch-original harmonic partial

$\tilde{\theta}_k$ phase for the *i*-th pitch-tuned harmonic partial

$\hat{\theta}_k^n$ unwrapped phase of $\tilde{\theta}_k^n$ versus $\hat{\theta}_k^{n-1}$

$\phi_k^n(t)$ cumulated phase for the *k*-th partial

REFERENCES

Chao, Y. R., 1968, *A Grammar of Spoken Chinese*, University of California Press, Berkeley, CA, USA.

Chen, T. C., and Chung, K. L., 2001, "A New Randomized Algorithm for Detecting Lines," *Real-Time Imaging*, Vol. 7, No. 6, pp. 473-481.

Dodge, C., and Jerse, T. A., 1997, *Computer Music: Synthesis, Composition, and Performance*, Schirmer Books, Prentice Hall International, NY, USA.

Faires, J. D., and Burden, R., 1998, *Numerical Methods*, Books/Cole Publishing Company, Pacific Grove, CA, USA.

Kim, Y., Kim, S. H., and Kwak, Y. K., 2005, "Dynamic Analysis of a Nonholonomic Two-Wheeled Inverted Pendulum Robot," *Journal of Intelligent and Robotic Systems: Theory and Applications*, Vol. 44, No. 1, pp. 25-46.

Kuroki, Y., 2006, "A Small Biped Entertainment Robot Creating Attractive Applications," *Springer Tracts in Advanced Robotics*, Vol. 24, pp. 13-20.

Li, Z. -C., 1990, *Computer Transformation of Digital Images and Patterns*, World Scientific, Teaneck, NJ, USA.

Nakamura, M., and Sawada, H., 2006, "Talking Robot and the Analysis of Autonomous Voice Acquisition," *IEEE International Conference on Intelligent Robots and Systems*, Beijing, China, Article No. 4059157, pp. 4684-4689.

NCVS, 2008, NCVS – National Center for Voice and Speech, Denver, Colorado, USA, Available: http://www.ncvs.org/ncvs/about/people/pavarobotti.htm

Otsu, N., 1979, "Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-8, pp. 62-66.

Stylianou, Y., 1996, "Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification," *Ph.D. Dissertation*, Ecole Nationale Supèrieure des Télécommunications, France.

Stylianou, Y., 2005, "Modeling Speech Based on Harmonic Plus Noise Models," *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 3445 LNAI, pp. 244-260.

Manuscript Received: Dec. 10, 2008
Revision Received: May 02, 2009
and Accepted: Jun. 02, 2009