

A face robot for autonomous simplified musical notation reading and singing

Chyi-Yeu Lin^{a,*}, Li-Chieh Cheng^a, Chang-Kuo Tseng^a, Hung-Yan Gu^b, Kuo-Liang Chung^b,
Chin-Shyurng Fahn^b, Kai-Jay Lu^b, Chih-Cheng Chang^c

^a Department of Mechanical Engineering, National Taiwan University of Science and Technology, 10607, Taipei, Taiwan

^b Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, 10607, Taipei, Taiwan

^c Institute of Automation and Control, National Taiwan University of Science and Technology, 10607, Taipei, Taiwan

ARTICLE INFO

Article history:

Received 17 January 2011

Received in revised form

12 May 2011

Accepted 2 July 2011

Available online 28 July 2011

Keywords:

Face robot

Facial expression

Musical note interpretation

Voice synthesis

ABSTRACT

This research is aimed to devise an anthropomorphic robotic head with a human-like face and a sheet of artificial skin that can read a randomly composed simplified musical notation and sing the corresponding content of the song once. The face robot is composed of an artificial facial skin that can express a number of facial expressions via motions driven by internal servo motors. Two cameras, each of them installed inside each eyeball of the face, provide vision capability for reading simplified musical notations. Computer vision techniques are subsequently used to interpret simplified musical notations and lyrics of their corresponding songs. Voice synthesis techniques are implemented to enable the face robot to sing songs by enunciating synthesized sounds. Mouth patterns of the face robot will be automatically changed to match the emotions corresponding to the lyrics of the songs. The experiments show that the face robot can successfully read and then accurately sing a song which is assigned discriminately.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, more and more intelligent robots have been developed for service and entertainment applications. Robots with various human-like facial expressions and conversation capabilities are useful for attracting attention from human beings and therefore are especially suitable for applications involving robot and human interaction.

Hara and Kobayashi [1] at Tokyo University of Science began researching in face robot field in 1990. Their face robot is anthropomorphized by a very realistic artificial facial skin and a set of deformation patterns used to create facial expressions. In that first face robot, multiple pneumatic muscles are placed behind its skull in order to pull at various points underneath its silicone rubber facial skin and create facial expressions. Its head is 1.2 times larger than a real human head. To shrink the head size to be similar to a human head, shape memory alloys are used as actuators in their revised version to replace the pneumatic muscles. Inside the robot head, many fans are installed for fast cooling to those shape memory alloys in order to speed up the restoration processes to their unexpanded statuses.

At MIT, Cynthia Breazeal developed an interactive face robot called Kismet [2,3]. There are a total of 15 DOFs on the face, including eyebrows, ears, eyeballs, eyelids, and mouth. To give

Kismet's vision and hearing abilities, each eyeball is installed a camera, and each ear is equipped with a microphone. Kismet can interpret emotions of the people in front of him and can show various facial expressions including happiness, anger, sadness, and surprise.

The Kobayashi Lab in Tokyo University of Science developed a female receptionist robot, called SAYA [4]. Her skin is made of a kind of silicone rubber. Many high speed pneumatic actuators are installed in the body and under the face to create head and arm motions as well as facial expressions. An air pump is installed in the lower body of her so that she becomes stationary and be limited to sit behind the desk. SAYA can track and interpret emotional statuses of the people in front of her and respond with proper facial expressions.

The Takanishi Laboratory in WASEDA University started to research on W.E. (Waseda Eye) series robots [5,6] in 1995, and a 59-DOF robot, WE-4R [7], was created in 2004. A large number of sensors are installed in WE-4R so that it can see, hear, smell, and feel the sense of touch and the sense of temperature. Many facial expressions, including those of anger, happiness, surprise, disgust, sadness, and fear, can be created by changing the shapes of exteriorly attached eyebrows, ears, eyeballs, eyelids, and lips. WE-4R is highly sensitive and sociable; he shows different emotional statuses and body motions in order to respond to people's actions toward him. Much earlier than this, the μ research group in the same university has developed WABOT-2 [8], an anthropomorphic robot that can play keyboard instruments. Its hands can tap softly on keys, legs can handle bass keys and the expression pedal, eyes

* Corresponding author. Tel.: +886 2 27376494; fax: +886 2 27301187.

E-mail address: jerrylin@mail.ntust.edu.tw (C.-Y. Lin).

can read a score, and mouth and ears can be used to converse with people.

Hirth et al. [9] presented a behavior-based emotional control architecture for an android head, ROMAN, in 2005. This architecture is based on 3 main parts: emotions, drives and actions which interact with each other to realize the human-like behavior of robots. Hanson Robotics [10] and Kokoro-Dreams [11] had produced highly realistic face robot products with full facial expression capabilities and extremely realistic face appearance. The facial skins of these robots are very close to those of human beings in terms of color, appearance, and shape. Hanson robotics had seven different face robot models for sale in 2007. In the same year, Kokoro-Dream had four face robot models. All of them are combined with a realistic human-like body with limited arm motion capabilities.

Although face robots have been made with realistic appearances, but there are associated with limited functions and applications. In this research, we aim to create a new entertainment function and an application for the face robot and thus open a new direction for development of entertainment robots. The face robot is made to autonomously sing songs as a show performance with a combination of musical notation recognition and voice synthesis techniques. This unique function has never been revealed or performed by any of the face robots in the literature. In the prior arts, some commercialized singer robots such as WowWee Alive Elvis® [12] can sing pre-programmed songs. This kind of singing robots cannot sing songs that are not stored in their internal storage devices. HRP humanoid robots [13,14] developed by AIST research team in Japan can follow human to sing the song after listen a small section of the prelude of the song sung by human. However, the performance capability of HRP humanoid robots in singing is also limited by their song data bases. The biggest difference and also the advantage of our robot singer from others is that there is no limitation on the number of songs that can be sung. In the proposed system, a human user can compose any song and corresponding lyrics, print it out in a certain format on a board, and show it to the robot singer. Then, the face robot singer will sing it after 30–40 s of time. This unique and innovative interaction capability enables the robot singer to entertain human audience.

In this research, we have done a 100 subject questionnaire survey to evaluate our face robot performance. The results of this evaluation reassure that the real time face robot singing as a performance is quite entertaining and acceptable.

2. The hardware of the face robot

2.1. System configuration

Designed to perform real time musical notation reading and singing, the face robot needs basic senses of hearing and vision and also speaking ability. In correspondence, microphone, camera, and speaker are equipped on the face robot. To enable a face robot to generate facial expressions, a number of actuators are equipped inside the skull to pull at the control points under the artificial face skin so as to deform it—creating facial expression. The material of the facial skin must be flexible and soft. Servo motors and Lynxmotion SSC32 servo controller are selected due to the compact size and simplicity. The SSC32 servo controller has 32 PWM output channels so that a single board can be used to control all servo motors. The system configuration of the face robot developed in this work can be seen in Fig. 1.

For easy maintenance and repair, all servo motors used inside the skull were divided into four modules. The top module including 8 motors was used to generate motions of the upper face. The eye module including 4 motors was used for generating eyeball motions. The chin module including 7 motors was used for

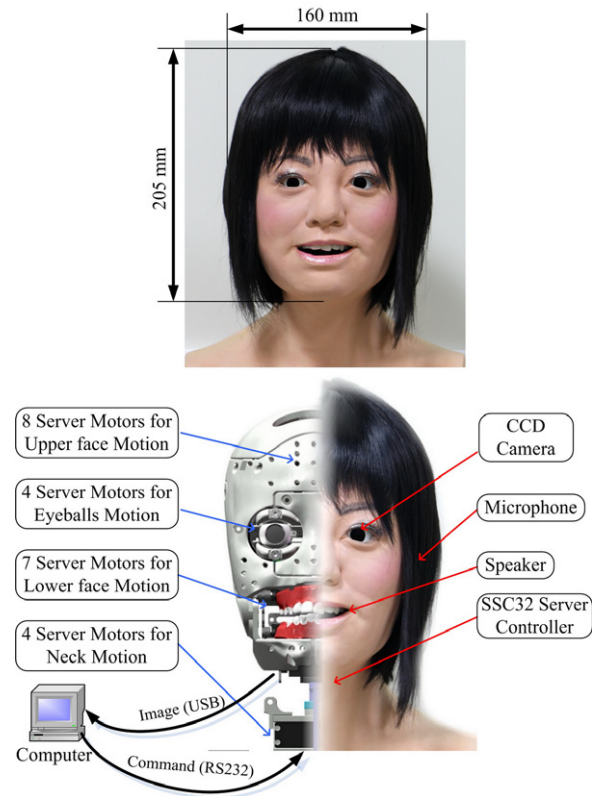


Fig. 1. System configuration of the face robot.

generating motions of the lower face. The neck module including 4 motors was used to generate motion of the neck. Rotational angle information for each servo motor corresponding to every facial expression is stored in a PC. When a specific facial expression is required display on the face robot, a set of pre-programmed commands is sent to a servo motor controller so as to create the expression. Commands are transmitted to the controller by the RS232 interface. USB interface transmits signals between two cameras and two microphones and the PC. An audio line transmits audio signals to a speaker inside the face robot to create vocal effects.

2.2. Artificial facial skin

The facial skin of the robot that can be deformed and show a number of facial expressions is commonly made of a type of silicone rubber with a proper color dye. This facial skin is designed to have various thicknesses in different regions to be similar to a human face skin. As shown in Fig. 2, the facial skin used in this work is based on a human model who is a 20 year-old woman. This type of technology is often applied to art and special effect in the cinema. Its procedure is a serial of various material molding processes. First of all, reproduce a human face that we want to mimic by molding with impression alginate. After alginate solidification, cover and fix it by pasting plaster bandages. Then, finish plaster, silicone, and clay molding in sequence. The purpose of molding clay mold is to make detail modification easier. After clay mold done, molding it with several layers of fiber reinforced plastic (FRP) material. Next, Inject polyurethane rubber [15], or silicone rubber [16] into the FRP mold. Finally, release the skin and finish the whole manufacture procedure by trimming, painting, and making up.

The number and locations of control points of the facial skin for deformation can affect the degree of reality of facial expressions. In this research, control point selection was based on suggestions

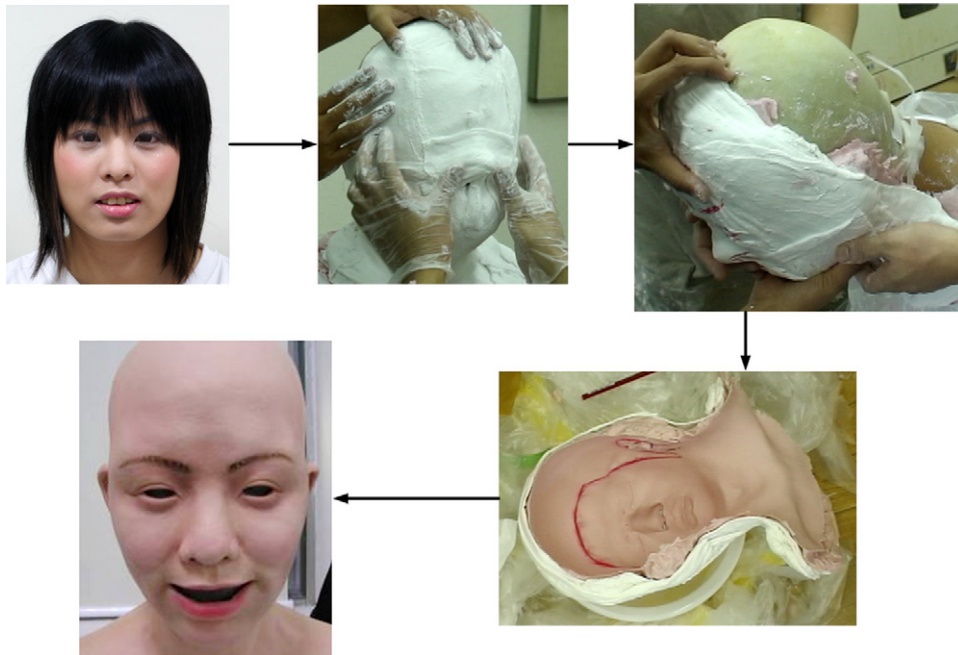


Fig. 2. Manufacture process of the artificial facial skin.

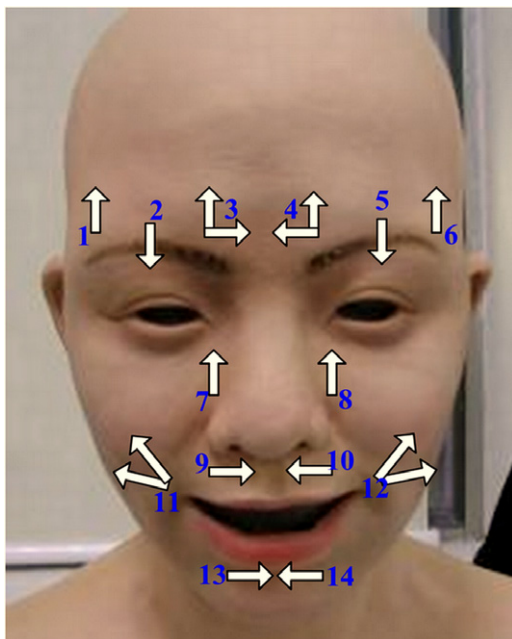


Fig. 3. Pulling points and directions for facial expressions.

proposed by Keith Waters [17] and also on the number of muscles required to create common facial expressions. Fig. 3 shows that android facial control points and their pulling directions. White arrows represent pulling directions of control points. Every control point is pulled by one or two steel wires with 0.4 mm diameter. The total number of control points on a face is 14. Points from no. 7 to no. 14 are related to speaking mouth pattern. All of them are on lower face. Point no. 9 and no. 13 are pulled by the same servo motor. Point no. 10 and no. 14 are pulled by another the same motor. Point no. 7 and no. 8 are pulled individually by two motors. Control points located at two corners of the mouth are shared by expression happy and expression sad so that point no. 11 and no. 12 are pulled by four motors in four directions.

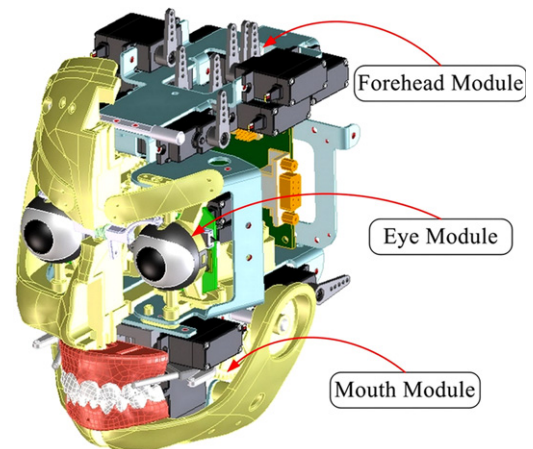


Fig. 4. Rapid prototyping skin support frame.

2.3. The skull mechanism

Since the silicone rubber facial skin is soft and stretchable, a support frame is needed to keep it in position, as shown in Fig. 4. The surface of the support frame is so complicated that it is suitable for using a rapid prototyping production. Many holes are included on the frame, through which wires attached to the servo motors can be connected to the interior surface of the face skin. For easy installment and maintenance, the frame is composed of three assemblies, the forehead module, the eye module, and the mouth module. The chin is included as an independent part of the mouth module so as to enable to open and close the mouth.

Each eyeball is equipped with a camera, and it has both vertical and horizontal degrees of freedom. The eyeball can be rotated horizontally to the left and right on each side by 45° . The eyeball is allowed to be lifted by 45° and turned down by 60° . The viewing angle of the camera is 45° , making the overall horizontal viewing angle 65° on both left and right sides and the overall vertical viewing angle 65° to the top and 80° to the bottom. Few existing face robots have tongues in expression applications. The face robot

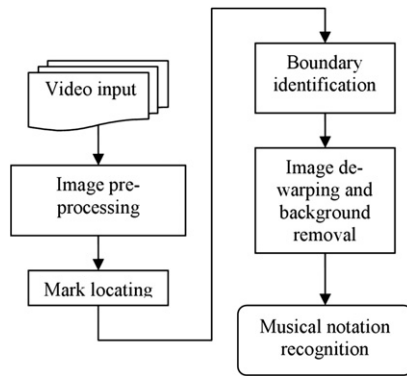


Fig. 5. Flowchart of our image capture system.

developed in this research is equipped with a tongue to assist in some consonant pronunciations in speaking.

3. Musical notation recognition

Vision equipment and techniques are used to interpret the musical notations. Two experimental webcams installed in the eyeball of the face robot are both Logitech QuickCam[®] Sphere MP, with video capture up to 640×480 pixels, still image capture of 1280×960 pixels, and frame rate of 30 frames per second. The focus of this kind of webcam is fixed. Although two such cameras are installed in the robot head, only one of them is used to capture the images of the simplified musical notations in front of them. Stereo vision capabilities by two cameras support the development of subsequent applications of the face robot which require distance information. Fig. 5 illustrates a flowchart of our image-capturing system. In addition to image pre-processing, the system is comprised of procedures for location marking, boundary identification, image de-warping, and background removal. The main blocks of the flowchart are elaborated in the following sections.

The RGB image captured by the webcam is transferred to a HSV image, from which the gray level image is extracted and employed in later steps. Based on the both hue and saturation level images, a color filter to segment the regions whose colors are similar to that of a prescribed mark. The regions of interest are regarded as a candidate locations of the mark applied in order to detect the paper of simplified musical notation placed on a stand.

In a natural scene, it is very difficult to locate simplified musical notation printed on paper. For finding the position of the paper more easily, we have added a simple mark which is easy to be recognized in a captured image. The mark that we have chosen is our school emblem, as shown in Fig. 6, and it is placed on the left-upper corner of the paper on which the simplified musical notation is printed. By means of color and geometry cues, we can effectively locate this mark in a complex background. In order to capture the complete image of the simplified musical notation, we must ensure all information appears fully in the field of vision. After acquiring the position of the mark from the image, the robot will move its neck to regulate the line of sight until the captured image of the simplified musical notation is also totally inside the field of vision.

The streaks of the paper edges in the captured image from the webcam are not very clear, especially when the paper is situated in front of a white wall. In order to make the edges more obvious, we glue the paper with simplified musical notation onto a black pasteboard. Once the captured image of the simplified musical notation is completely presented on the screen by properly locating the mark, we start to detect the edges of the paper from the neighborhood of the mark using a clockwise tracking sequence. In the following, we simulate the boundary identification algorithm



Fig. 6. Picture of the NTUST school emblem.

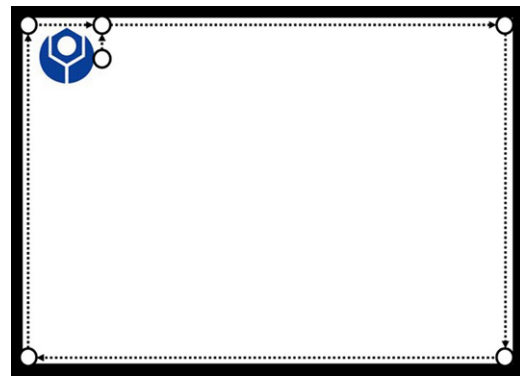


Fig. 7. Illustration of the tracking sequence of boundary identification.

with a small circle rolling from the right side of the mark. The circle's rolling path is illustrated in Fig. 7, where the circle rolls away from the center of the paper until it touches the boundary. Here, the boundary of the paper is identified by the Sobel edge detector [18]. Next, we record the path and estimate the extent of the paper appears in the field of vision.

After understanding the range of the simplified musical notation illustrated, we preserve those pixels inside the range and exploit four corner points of the paper to constitute a skewed quadrilateral image used for calculation of de-warping parameters [19]. According to these, the skewed quadrilateral image can be restored to a near rectangular one. At the same time, we also apply an image interpolation method to increase the resolution. As a result, an upright and high-definition image of simplified musical notation in a simple background is produced for optical character recognition.

The realization of our robot vision system takes place in a developing environment including Borland C++ Builder 6 and Microsoft Windows XP Professional; the host computer is equipped with a Pentium Mobile 1.8 GHz CPU and 1 GB DRAM. In the beginning of the image capture, the paper with simplified musical notation is laid on a stand in front of the robot's head at a distance of about 50 cm. The glancing image captured by the webcam is usually quite complex, as shown in Fig. 8(a), where part of the image near the bottom is occluded by the eyelid of the robot's face. After carrying out the entire capture procedure described previously, we can obtain an upright image of simplified musical notation, without background, with resolution of 2100×1530 pixels, as demonstrated in Fig. 8(b). The total execution time for generating such an image is about two seconds, exclusive of the I/O access time, motor rolling time, mark locating time, and so on.

After performing image de-warping and background removal, a simplified musical notation recognition technique is developed to

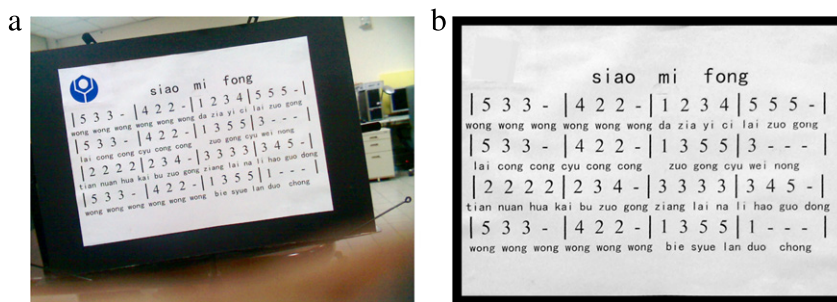


Fig. 8. High-definition image generation: (a) glancing image captured by the webcam; (b) processed upright image.

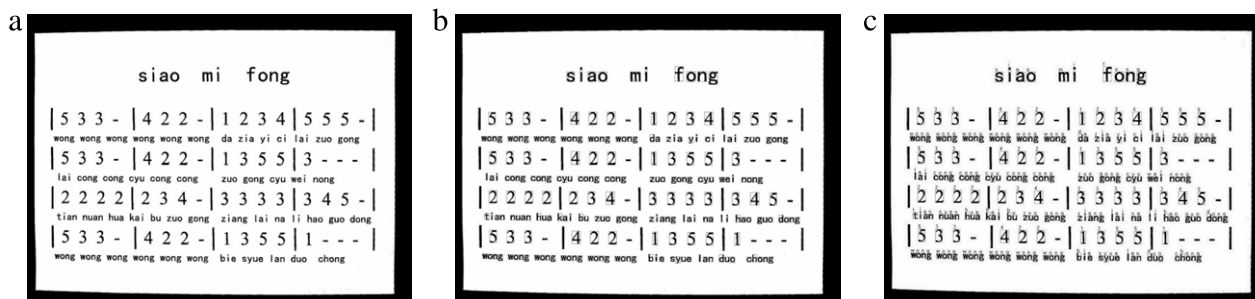


Fig. 9. Illustration of the simplified musical notation recognition technique: (a) the binary image; (b) the partitioned blocks; (c) the recognized result.

interpret the simplified musical notation image. The recognition process consists of four steps. In Step 1, the de-warped image without background is binarized by using Otsu's thresholding method [20]. The resultant image is shown in Fig. 9(a). In Step 2, the binarized image is partitioned into non-overlapped blocks, where each block contains a simplified musical notation or an English character, by using projection techniques corresponding to the *x*-axis and *y*-axis, respectively. These partitioned blocks are shown in Fig. 9(b). In Step 3, for each block, to find the matched value between the simplified musical notation and the model block, we find the affine-invariant matched value between the block containing the simplified musical notation and the model block stored in the database in advance. The matched value is used to recognize the meaning of the simplified musical notation in the block, as shown in Fig. 9(c). Finally, the recognized English characters into meaningful words are merged according to the distance between two adjacent English characters. The proposed simplified musical notation recognition technique provides high recognition accuracy for the input simplified musical notation image. The recognition output, which includes English words and simplified musical notations, will be used as the input of the Mandarin singing voice synthesis system.

4. Mandarin voice synthesis

To synthesize music signals, additive synthesis, subtractive synthesis, and FM (Frequency Modulation) synthesis are notable techniques [21,22]. In this paper, however, the technique of HNM (harmonic plus noise model) originally proposed by Stylianou [23, 24] is used as a foundation and then extended. We selected HNM so that the synthetic Mandarin singing voice would have a higher signal clarity and naturalness level. HNM splits the spectrum of a signal frame into two halves of unequal widths to better model the spectrum. The lower frequency half is modeled as harmonic partials, while the higher frequency half is modeled as noise signal components.

Here, we adopt the syllable as the unit for synthesis processing. This is because Mandarin is a syllable-prominent language, and each syllable is of the structure CxVCn. The Cx of a syllable may

be null, a voiced consonant, or an unvoiced consonant, while the Cn may be null, or nasal, as in /n/ or /ng/. Additionally, the nucleus, V, may be a vowel, a diphthong, or a triphthong. If the Cx is a long unvoiced consonant (e.g., /s, p/), its synthetic signal will be generated as a noise signal with HNM. If the Cx part is a short unvoiced consonant (e.g., /b, d/), its synthetic signal will be directly copied from the corresponding part in the recorded syllable. Otherwise, the Cx is a voiced consonant (e.g., /m, r/) and is considered together with the remaining phonemes. Then, their synthetic signals are generated as harmonic partials plus noise signal with HNM.

4.1. Score file parsing and interpretation

The data of a song score is stored as a text file and consists of a sequence of lines. Each line contains one note of information, i.e., pitch, beats, and lyrics, except for the first line. The information stored in the first line consists of song name, tempo (e.g., 120 means 120 beats per minute), and fullness ratio (e.g., 80 means only 80% of a note's duration is used).

The pitch of a note is represented by a symbol in the first field of a line. By interpretation, the symbol is converted to a numeric value of pitch frequency. After all notes' pitch frequencies are determined, automatic key-shifting is performed. This is done so that the pitch range of the score file is matched to the pitch range of the person who utters the Mandarin syllables for analyzing HNM parameters. The second field of a line contains the number of beats. Hence, the duration of a note can be computed by multiplying it with the beat length, i.e., 60 over the tempo value. However, the duration of a note is usually not fully sung because some small ratio of this duration is reserved for breathing and transitioning to the following note. The third field of a line contains the lyrics of a note. Each note usually has a unique lyric (i.e., a syllable) assigned to it. However, it may occur that two or three consecutive notes are assigned to a same lyric. Then, portamento singing, i.e. pitch gliding or glissando, must be synthesized for the lyric syllable that has two or more notes assigned to it. In the score file, this situation is hinted with a convention that if the third field of a note is placed the special character, "|", this note is assigned the same lyric as its preceding note.

4.2. Signal waveform synthesis

When applying HNM to synthesize signal samples for a lyric syllable, it is found some issues that are not discussed in the literature on HNM. The first issue is how to estimate the spectral envelope in order to keep the timbre of synthetic syllable signals consistent. Note that we only intend to record each syllable's utterance once, and then we adjust the recorded syllable's pitch frequency to the target frequency of a note having this syllable assigned as its lyric. When the target frequency of a syllable is known, the values of this syllable's HNM parameters must be adjusted in a way that the timbre can be kept consistent. In addition, the second issue is how to warp the time axis of a synthetic syllable so that a more fluent syllable signal can be synthesized. Note that a simple time-warping method, i.e., linear warping, will usually result in lower perceived fluency when a syllable's duration is lengthened or shortened.

4.2.1. Phoneme duration planning

When a syllable starts with a short-unvoiced phoneme, e.g., /bau/, the time length of the short-unvoiced is planned as the corresponding phoneme length in the recorded syllable. When a syllable starts with a long-unvoiced phoneme, the length of the long-unvoiced is planned by multiplying its original length by a factor L_u . L_u is computed as the synthetic syllable's length divided by the recorded syllable's length. However, the value of L_u is confined within the range 0.6–1.4. The values, 0.6 and 1.4, are determined empirically. They are used to reserve a minimum duration to have the phoneme perceived, and limit the maximum duration to simulate a real person's uttering.

For the voiced phonemes of a syllable, the phoneme durations are planned according to an observation. That is, the consonant-to-vowel duration ratio will become smaller when the syllable is uttered within a sentence instead of in isolation. Consider the example syllable, /man/. Suppose that in the recorded signal of /man/, the three phonemes, /m/, /a/, and /n/, occupy R_m , R_a , and R_n seconds, respectively, and $R_v = R_m + R_a + R_n$. Also, suppose that D_m , D_a , and D_n represent the time lengths of the three phonemes within the synthetic syllable, and $D_v = D_m + D_a + D_n$. The value of D_m is planned by multiplying a duration reduction rate, r (initially set to 0.85), with the time ratio (R_m/R_v) of its counterpart, R_m , in the recorded syllable. In the same way, the value of D_n is planned. By trying to decrease the value of r iteratively, the values of D_m and D_n are decreased gradually, and the value of D_a finally becomes sufficiently large (i.e., $D_a > D_v * 0.5$). After the values of the durations are determined, a mapping function from the phonemes in the synthetic syllable to the corresponding phonemes in the recorded syllable can then be established. This mapping function is a piecewise linear time-warping function as illustrated in Fig. 10.

4.2.2. Pitch-contour generation for portamento singing

When a syllable is assigned more than one note, it should be sung in a portamento manner. That is, the pitch-contour of the syllable should be smoothly transitioned from the former note's pitch to the latter note's pitch in the middle portion. An example pitch-contour is shown in Fig. 11. The duration of the syllable is divided into three time intervals. The left and right intervals are planned to sing stable pitches of the two notes in order that they can be explicitly perceived. And the middle interval is used to transit the pitch smoothly.

Suppose that the two notes to be sung in portamento are of the pitch frequencies P_a and P_b . The control points are divided within the voiced part of the synthetic syllable into three groups. Then, the control points within the first and third groups are directly assigned the pitches of P_a and P_b respectively. But for the n -th

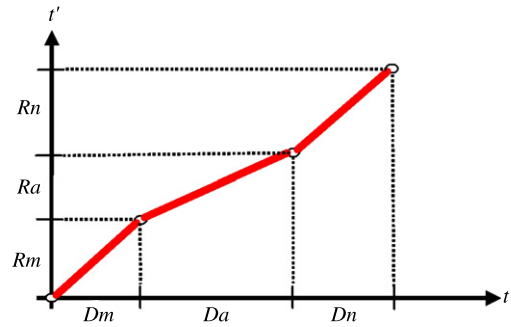


Fig. 10. Piecewise linear mapping function.

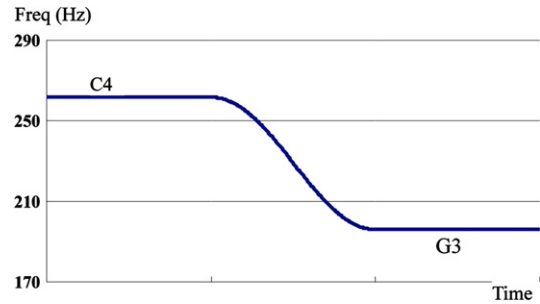


Fig. 11. An example pitch-contour for a portamento singing.

control point in the second group, its pitch, P^n , is defined here with a cosine based function. That is,

$$P^n = \frac{(P_a + P_b)}{2} + \frac{(P_a - P_b)}{2} \cos\left(\frac{n}{M}\pi\right) \quad (1)$$

where M is the number of control points in the second group. This cosine based transition function has a good characteristic that the slopes at left and right ends are both zero and the transition curve hence connects smoothly to the left and right stable pitch segments.

4.2.3. Pitch-tuned HNM parameters

According to the constructed mapping function in Fig. 10, each analysis frame's time position on a recorded syllable's time axis can then be mapped to a time position on a synthetic syllable's time axis. The mapped time position on a synthetic syllable is also called a control point. Therefore, on a control point, the pitch-original HNM parameters, A_i (amplitude), F_i (frequency), and θ_i (phase), for the i -th harmonic partial can be obtained by referring to its corresponding analysis frame. However, the parameters \tilde{A}_k , \tilde{F}_k , and $\tilde{\theta}_k$, for the pitch-tuned harmonic partial should be determined carefully in order to keep timbre consistent.

To have consistent timbre, the spectral envelope must be unchanged. Therefore, the amplitude \tilde{A}_k of the pitch-tuned harmonic partial located at frequency \tilde{F}_k should be interpolated according to the original spectral envelope that is defined by the sequence of pairs (F_i, A_i) . Here, third-order Lagrange interpolation [25] is used. In more detailed, we first find an original harmonic frequency, F_j , from the sequence, F_1, F_2, F_3, \dots , that is nearest to and less than \tilde{F}_k . Then, the four original partials of the frequencies, F_{j-1}, F_j, F_{j+1} , and F_{j+2} , are used to perform third-order Lagrange interpolation to compute the value of \tilde{A}_k . Similarly, the phase $\tilde{\theta}_k$ of the pitch-tuned harmonic partial located at frequency \tilde{F}_k can be interpolated as well. However, the phases of the original partials must be unwrapped beforehand to prevent phase discontinuities. As to the order of interpolation, the sound synthesized by second-order interpolation is perceived to be less delicate. Therefore, third-order interpolation is selected.

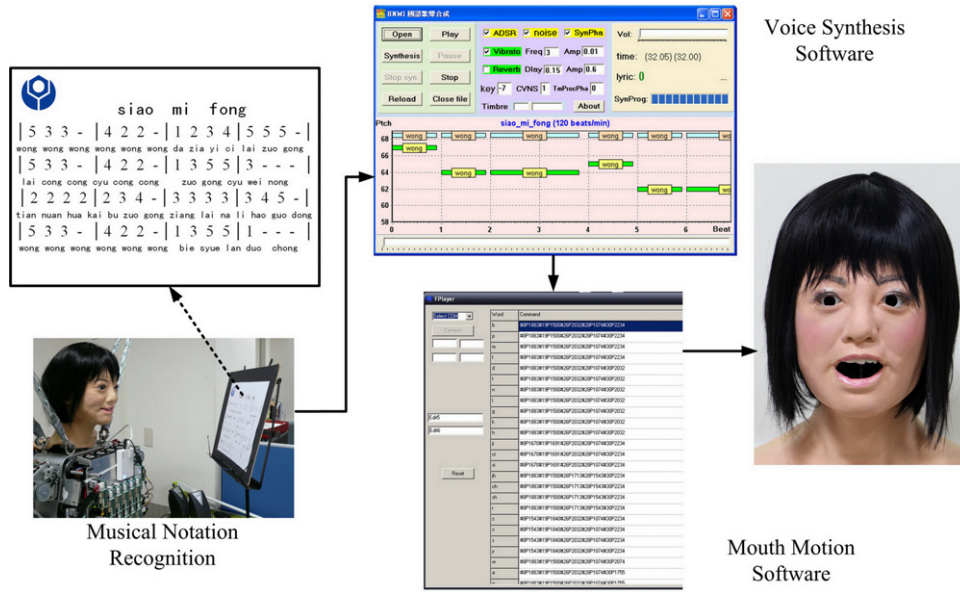


Fig. 12. The autonomous read-and-sing integration demonstration of the face robot.

4.2.4. Signal-sample synthesis

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

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

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

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

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

where L is the number of harmonic partials, T^n is the number of sample points between the n -th and $(n + 1)$ -th control points, 22050 (Hz) is the sampling rate, $a_k^n(t)$ is the time-varying amplitude of the k -th partial at time t , $\phi_k^n(t)$ is the cumulated phase for the k -th partial, $f_k^n(t)$ is the time-varying frequency (in Hz) 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 $\hat{\theta}_k^{n-1}$. In Eqs. (3) and (5), linear interpolation is used.

The noise signal is synthesized here as a summation of sinusoidal signal components. Let G_k be the frequency of the k -th sinusoidal component. Here, G_k is defined as $100 \cdot k$ (Hz) according to the thesis of Stylianou. For the n -th control point, the index k of G_k is not started at 1, and its starting value is determined according to the MVF (maximum voiced frequency) value of this control point, i.e., $\lceil \text{MVF}(n)/100 \rceil$. The MVF value of an analysis frame is determined during HNM parameter analysis.

Let B_k^n 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 are inversely Fourier-transformed to the spectral domain. Then, in terms of the spectral magnitude coefficients, X_j , $j = 0, 1, \dots, 2047$ (after exponentiation), B_k^n can be obtained by linearly interpolating the two adjacent magnitudes, X_i and X_{i+1} , whose frequencies (indexed by i) surround the frequency of G_k .

5. Matching mouth patterns with voice content

The integration demonstration of the simplified musical notation reading and singing face robot is shown in Fig. 12.

After the face robot interprets the simplified musical notation and synthesizes the song content, the mouth patterns to match the voice content must be implemented so as to perform the song while singing in a natural manner. Chinese pronunciation consists of single syllables exclusively. Each word is composed of a vowel, a consonant, or the combination of a vowel and a consonant, and there are a total of 23 vowels and 37 consonants. The voice synthesis module determines the lyrics of a song for the face robot to sing. The lyrics of a song will be used to determine the corresponding mouth pattern when a sound is to be created. The mouth pattern generator will first check if the voice content consists of only a vowel or a consonant, or a combination of them. For a combination of a vowel and a consonant, the mouth patterns need to be changed quickly from the first part, such as the consonant “n”, as shown in the left photo of Fig. 13, to the vowel “a”, as shown in the right photo of Fig. 13. For a voice content containing only a vowel or a consonant, the mouth pattern remains unchanged during the length of the sound.

In human vocal language, a phoneme is the smallest posited structural unit that distinguishes meanings. As separate words are fed into the mouth pattern software, they will be split into the required phonemes for speech synthesis, as shown in Table 1. Different phonemes will produce different pronunciations, but for many of the pronunciations, their mouth shapes are quite similar, so the mouth actions need to be redefined with 12 different mouth shapes [26]. For a separate word like “wang”, the word pronunciation is broken down into phonemes in the form of “w-ang”, which are then to be matched with respective control command for corresponding mouth shapes. Thereafter, these commands are converted to signals that control the movement of the mouth on the face robot during singing or voice broadcasting. The mouth action between every separate word is set to “silence”, such that all mouth actions can be linked to form a continuous speech delivery. The majority of mouth actions involve the opening and closing of the chin. With regard to the voice output speed, the maximum chin opening angle is 30° when the mouth is wide open, and the fastest rotation speed of the motor is 0.562 s. For the voice broadcast, the largest chin opening angle when talking is set to be 15° , and the time it takes to reach this stage is about 0.281 s. The talking speed of the face robot is about 3.5 syllables per second at the maximum.

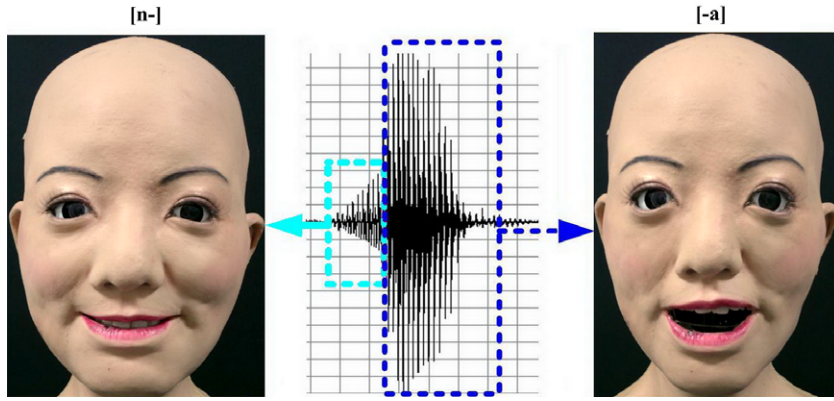


Fig. 13. Various mouth patterns: (left) mouth pattern for vowel “n-”; (right) mouth pattern for consonant “-a”.

Table 1
Chinese phonemes and mouth patterns.

Chinese phoneme	Example	Robotic mouth pattern	Human mouth pattern
b, p, m,	<u>big</u>		
w, -ao, -ou, -ong, -iao, -iu, -u, -un, -yu, -yun, -yong	<u>go</u>		
r, -er,	<u>red</u>		
f	<u>fork</u>		
t, n	<u>thin</u>		
l	<u>lid</u>		
d, z, c, s,	<u>sit</u>		
ji, ci, si, jh, ch, sh,	<u>she</u>		
g, k, h, y, -i, -in, -ing,	<u>yard</u>		
-e, -ei, -an, -ang, -ie, -ian, -iang, -ui, -uan, -yue, -yuan,	<u>ate</u>		
-a, -ai, -ia, -ua, -uai, -uang,	<u>cat</u>		
-o, -ou, -en, -eng, -uo,	<u>toy</u>		

6. Evaluation and results

To scientifically analyze the response of human being when they watch the real time musical notation reading and singing performance of our face robot, an evaluation experiment with questionnaires was made. One hundred randomly selected subjects were interviewed on the street of Taipei and questionnaires were answered. The sampling distributions sorted by gender, age, and education background are shown in Table 2. Every subject was asked to watch a one-minute video first, which contains the performance of the real time musical notion recognition and singing

by our face robot. After that, every subject was asked to finish the questionnaire we designed. There are six groups of questions in the questionnaire. The statistic results of this survey are shown in Tables 3–8. For the facial expression evaluation, the effective questionnaires are 89 copies (89%) and the ineffective questionnaires are 11 copies (11%). For the singing performance evaluation, the effective questionnaires are 99 copies (99%) and the ineffective questionnaire is only 1 copy (1%).

The result of the survey on the questions of every group implies different meanings. In Table 3, three typical facial expressions, happiness, shock and anger, were involved in the facial expression

Table 2
The sampling distributions sorted by gender, age, and educated background.

Total	Gender		Age					Educated background		
	Female	Male	5–14 years old	15–24 years old	25–39 years old	40–64 years old	>65 years old	Under senior high school	Science and engineering related	Non science and engineering related
100	52	48	5	38	30	21	6	30	21	49

Table 3
Scores on the facial expressional reality of the face robot.




Item	1	2	3
Expression			
	Happiness	Shock	Anger
Correct recognition rate	85.39%	93.26%	65.17%
Average Score (0–10)	6.92	7.37	6.07

Table 4
Opinions on autonomous musical notation reading and singing of robot.

Item	Question	Opinion rate				
1	Have you ever seen any face robot performing musical notation reading and singing before?	Yes	No			
		12.12%	87.88%			
2	How do you think this technique after watching the face robot performance video?	Very interesting	Interesting	Unknown	Boring	Very boring
		26.26%	62.63%	10.10%	1.01%	0%
3	Do you agree robotic singing is entertaining?	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
		17.17%	50.51%	22.22%	8.08%	2.02%

Table 5
Evaluation on the vocal performance of this face robot singing.

Item	Question	Opinion rate				
1	Can you know what song is sung by the face robot in the video?	Yes	No			
		88.89%	11.11%			
2	Can you understand the lyrics by hearing this face robot singing?	Yes	No			
		68.69%	31.31%			
3	In terms of the voice only, do you know the singing sound is coming from a robot?	Yes	No			
		55.56%	44.44%			
4	How do you feel about the voice from the face robot?	It is human voice	Similar to human voice	Unknown	Not realistic, but acceptable	Not realistic and unacceptable
		6.06%	41.41%	11.12%	41.41%	0%

Table 6
Evaluation on the appearance of this face robot.

Item	Question	Opinion rate				
1	Looking at the face only, do you know that singer is a robot?	Yes	No			
		87.88%	12.12%			
2	How do you feel about the appearance of the face robot?	It is a real human	Similar to a human	Unknown	Easy to tell it is a robot	It is a robot
		1.01%	53.54%	5.05%	37.37%	3.03%
3	Do you agree the reality of the face robot appearance is more important than its singing ability?	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
		19.19%	36.36%	22.23%	20.20%	2.02%

Table 7
Evaluation on the overall presentation and performance of this face robot.

Item	Question	Opinion rate				
1	How do you think the overall presentation and performance of this face robot?	Excellent	Good	Regular	Bad	Awful
		3.06%	70.41%	23.47%	3.06%	0%

recognition test. The shocked face is the easiest one to recognize correctly, while the angry face is the hardest to be recognized correctly. All of the three facial expressions get an average score

over 6.0 in resemblance from the subjects who recognized them correctly. The linear resemblance score scale ranges from 0 to 10 while 0 meaning 0% of resemblance and 10 meaning 100%

Table 8
Opinions on the development of robotic entertainment industry.

Item	Question	Opinion rate				
1	Do you agree it is possible to use robot to do entertainment performances?	Strongly agree 15.15%	Agree 60.61%	Undecided 20.20%	Disagree 4.04%	Strongly disagree 0%
2	Do you agree that robotic performance will be a new entertainment form in the future?	Strongly agree 16.16%	Agree 59.60%	Undecided 24.24%	Disagree 0%	Strongly disagree 0%
3	Will you stand for related organizations should keep on developing the entertainment robot?		Yes 91.92%		No 8.08%	
4	Would you like to watch any robotic show if you have a chance in the future?		Yes 84.85%		No 15.15%	
5	Would you like to buy a ticket for watching any robotic show if you have a chance in the future?		Yes 62.63%		No 37.37%	

resemblance. It means the reality of the facial expressions shown on the face of our singer robot is accepted on average by people, but there is still quite large a space to improve. In Table 4, it is noted that the performance of autonomous musical notation reading and singing by the face robot is largely agreed to be new, interesting, and entertaining. Table 5 shows the face robot can sing the song in a recognizable standard, and most subjects accept the synthesized singing voice of the face robot. In Table 6, the opinions tell that the resemblance of the appearance of the face robot is important for the entertaining performance of real time singing to be deeply appreciated. In Table 7, the majority affirms the overall presentation and performance of the face robot is in a good standard. In Table 8, most subjects agree that the robot entertainment is worth further developing, and they stand for that related research organizations should keep on developing entertainment robots. Moreover, most of them would like to watch any robotic show either free or buying a ticket.

These results strongly support our idea to develop entertainment robots and related applications. It shows that people love robots and they like to watch performances of these robots. Our face robot for autonomous musical notation reading and singing is the first step and a milestone for developing entertaining robots leading to future robot entertainment industry.

7. Conclusions

The face robot developed in this research is anthropomorphized by a unique integrated function of being able to read simplified musical notation and then sing the song content with corresponding mouth patterns. The face robot equipped with 24 servo motors can show a number of realistic facial expressions and speak like a human.

An automatic image-capturing method is created to generate a high-definition upright image of simplified musical notation. Our robust vision system can automatically detect simplified musical notation when it is placed in front of the face robot, and it can properly capture the image of numbered musical notation using a webcam. Even if the original image captured by the webcam is skewed and partially covered by the eyelid, the system can still de-warp the skewed image, remove the background, and raise the resolution about 2.5 times in a naturally complex scene to ease the subsequent simplified musical notation interpretation.

This paper described the following procedure: an inputted Mandarin song score is parsed to extract each note's information firstly. Next, the lyric associated with a note is used to load a corresponding HNM parameter values. In terms of the parameter values, signal samples for a lyric can then be synthesized with the HNM-based, extended method proposed here. In the following subsections, the details of the processing steps were described. The subsystem can synthesize a Mandarin singing voice in real-time, and the synthetic signal is very clear and natural.

The successfully integrated simplified musical notation reading and singing ability of the face robot in this work demonstrates

a promising direction for robot entertainment in the future. To increase the functionalities of intelligent robots in the entertainment sector, and more specifically in robot musical and speech talent applications, techniques to create more natural facial expressions and easier text-to-speech capabilities with highly accurate mouth pattern generation need continual investigation.

Acknowledgments

This research was financially funded by the National Science Council of the Republic of China (Taiwan) under grant numbers NSC 94-2212-E-011-032, NSC 94-2218-E-011-012, and NSC 95-2218-E-011-009. Their support made this research and the outcome possible.

References

- [1] F. Hara, H. Kobayashi, Face robot able to recognize and produce facial expression, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 1996, pp. 1600–1607.
- [2] C. Breazeal, B. Scassellati, How to build robots that make friends and influence people, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 1999, pp. 858–863.
- [3] C. Breazeal, B. Scassellati, Infant-like social interactions between a robot and a human caretaker, *Journal of Adaptive Behavior* 8 (2000) 49–74.
- [4] T. Hashimoto, M. Senda, T. Shiiba, H. Kobayashi, Development of the interactive receptionist system by the face Robot, in: Proceedings of the SICE Annual Conference, 2004, pp. 1563–1567.
- [5] A. Takanishi, S. Ishimoto, T. Matsuno, Development of an anthropomorphic head-eye system for robot and human communication, in: Proceedings of the IEEE International Workshop on Robot and Human Communication, 1995, pp. 77–82.
- [6] Takanishi Laboratory, <http://www.takanishi.mech.waseda.ac.jp/top/index.htm> (accessed 25.04.11).
- [7] H. Miwa, K. Itoh, M. Matsumoto, M. Zecca, H. Takanobu, S. Roccella, M.C. Carrozza, P. Dario, A. Takanishi, Effective emotional expressions with emotion expression humanoid robot WE-4RII, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004, pp. 2203–2208.
- [8] I. kato, S. Ohteru, K. Shirai, S. Narita, S. Suqano, T. Matsushima, T. Kobayashi, E. Fujisawa, The robot musician 'WABOT-2', *Robotics Amsterdam* 3 (2) (1987) 143–155.
- [9] J. Hirth, N. Schmitz, K. Berns, Emotional architecture for the humanoid robot head ROMAN, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2007, pp. 2150–2155.
- [10] Hanson Robotics, <http://hansonrobotics.wordpress.com/> (accessed 25.04.11).
- [11] Kokoro Company Ltd., <http://www.kokoro-dreams.co.jp/english/index.html> (accessed 25.04.11).
- [12] WowWee™ Group Limited, <http://www.wowwee.com/en/support/elvis> (accessed 25.04.11).
- [13] T. Nakano, M. Goto, VocalListener: a singing-to-singing synthesis system based on iterative parameter estimation, in: Proceedings of the 6th Sound and Music Computing Conference, 2009, pp. 343–348.
- [14] T. Otsuka, K. Murata, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, H.G. Okuno, Incremental polyphonic audio to score alignment using beat tracking for singer robots, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, pp. 2289–2296.
- [15] T. Hashimoto, S. Hiramatsu, H. Kobayashi, Development of face robot for emotional communication between human and robot, in: Proceedings of the IEEE International Conference on Mechatronics and Automation, 2006, pp. 25–30.

- [16] M. Hashimoto, C. Yokogawa, T. Sadoyama, Development and control of a face robot imitating human muscular structures, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 1855–1860.
- [17] K. Waters, A muscle model for animation three-dimensional facial expression, *Computer Graphics* 21 (4) (1987) 17–24.
- [18] R.C. Gonzalez, R.E. Woods, *Digital Image Processing*, 2nd ed., Prentice-Hall, New Jersey, 2002.
- [19] Z.C. Li, T.D. Bui, Y.Y. Tang, C.Y. Suen, *Computer Transformation of Digital Images and Patterns*, World Scientific, Hackensack, NJ, 1989.
- [20] N. Ostu, A threshold selection method from gray level histogram, *IEEE Transactions on Systems, Man and Cybernetics SMC-9* 1 (1979) 62–66.
- [21] F.R. Moore, *Elements of Computer Music*, Prentice-Hall, USA, 1990.
- [22] C. Dodge, T.A. Jerse, *Computer Music: Synthesis, Composition, and Performance*, 2nd ed., Schirmer Books, New York, 1997.
- [23] Y. Stylianou, Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification, Ph.D. Thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, 1996.
- [24] Y. Stylianou, Modeling speech based on harmonic plus noise models, *Nonlinear Speech Modeling and Applications* 3445 (2005) 244–260.
- [25] J.D. Faires, R. Burden, *Numerical Methods*, 2nd ed., Books/Cole Publishing Company, Pacific Grove, CA, USA, 1998.
- [26] B. Tiddeman, D. Perrett, Prototyping and transforming visemes for animated speech, in: Proceedings of the Computer Animation, 2002, pp. 248–251.



Chyi-Yeu Lin was born in Taiwan in 1957. He received the Ph.D. Degree in engineering mechanics from the University of Florida, USA, in 1991. He is currently the chairman of the faculty in the Department of Mechanical Engineering and the director of Center for Intelligent Robots in National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. His research has been focused on intelligent robots, intelligent machines, computer vision, artificial intelligence, optimal design, and structural optimization.



Li-Chieh Cheng was born in Taipei, Taiwan, in 1973. He received the M.S. Degree in mechanical engineering from National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 2000. He is currently working toward the Ph.D. Degree in mechanical engineering in National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. His research has been focused on android mechanical design, human facial expression analysis, human face mimetic engineering, and artificial skin manufacturing.



Chang-Kuo Tseng was born in Taipei, Taiwan, in 1979. He received the M.S. and Ph.D. Degrees in mechanical engineering from National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2004 and 2009, respectively. His research has been focused on mechatronics and humanoid robots.



hiding.

Hung-Yan Gu received the B.S. and M.S. Degrees both in computer engineering from National Chao-Tung University (NCTU), Taiwan, in 1983 and 1985, respectively and received the Ph.D. Degree in computer science and information engineering from National Taiwan University (NTU), Taipei, Taiwan, in 1990. He is currently an associate professor in the Department of Computer Science and Information Engineering in National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. His research has been focused on speech signal processing, computer music synthesis, data compression, and information



Kuo-Liang Chung received the B.S., M.S., and Ph.D. Degrees in computer science and information engineering from National Taiwan University (NTU), Taipei, Taiwan, in 1982, 1984, and 1990, respectively. He is currently a professor in the Department of Computer Science and Information Engineering in National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. His research has been focused on image/video compression and image/video processing.

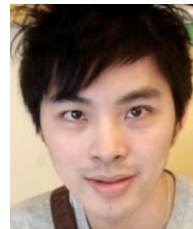


intelligence, and robot automatic control.

Chin-Shyurng Fahn was born in Tainan (Taiwan) in 1958. He received the B.S. Degree in electronic engineering from National Taiwan Ocean University (NTOU), Keelung, Taiwan, in 1981, and the M.S. and Ph.D. Degrees both in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1983 and 1989, respectively. He now serves as an associate professor in the Department of Computer Science and Information Engineering in National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. His research has been focused on robot vision, robot perception, robot



Kai-Jay Lu was born in Taichung, Taiwan, in 1982. He received the B.S. Degree in computer and communication engineering from Ming Chuan University (MCU), Taoyuan, Taiwan, in 2006. He received the M.S. Degree in computer science and information engineering from National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2009. His research interests include image processing, pattern recognition, robot vision, and robot automatic control.



Chih-Cheng Chang received the B.S. Degree in electronic engineering from Fu Jen Catholic University (FJU), Taipei, Taiwan, in 2006. He received the M.S. Degree in automation and control engineering from National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2008. His research interests include image processing and algorithms.