2011 IEEE-RAS International Conference on Humanoid Robots
Bled, Slovenia, October 26-28, 2011

"Mask-bot": a life-size robot head using talking head animation for human-robot communication

Takaaki Kuratate*†, Yosuke Matsusaka‡, Brennand Pierce*, Gordon Cheng*
*Institute for Cognitive Systems, Technical University Munich, Germany, Email: {kuratate, bren, gordon}@tum.de
†MARCS Auditory Laboratories, University of Western Sydney, Australia, Email: t.kuratate@uws.edu.au
‡National Institute of Advanced Industrial Science and Technology, Japan, Email: yosuke.matsusaka@aist.go.jp

***Abstract*—In this paper, we introduce our life-size talking head robotic system, "Mask-bot", developed as a platform to support and accelerate human-robot communication research. The "Mask-bot" hardware consists of a semi-transparent plain mask, a portable LED projector with a fish-eye conversion lens mounted behind the mask, a pan-tilt unit and a mounting base. The hardware is driven by a software animation engine controlling face model selection and calibration augmented by an OpenHRI-based speech communication module that generates text-to-speech responses for conversational exchanges. The system projects the calibrated animated head onto the mask. Via this process the head becomes a life-size talking head in real space as opposed to 2D flat screen space or stereo pseudo-3D screen space. "Mask-bot" can easily change the appearance and behavior of the face model, affording the means for easily evaluating new face platforms for a variety of characteristics, including AV speech synthesis and perception, without building a new actual robotic head, thus accelerating the design cycle for robot heads.**

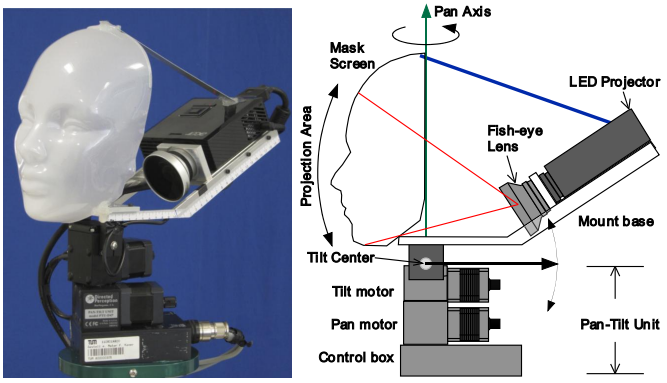

Fig. 1. *Mask-bot display system overview and structure diagram*

I. Introduction

The introduction of humanoid robots in to our daily lives, ranging in size from small desktop robots to adult-sized robots, is becoming increasingly common. To some, the main goal in developing such robots is in making them as realistic as humans. This means identifying and solving the many, often tough challenges associated with creating a realistic physical entity. Among the various attempts to build realistic face robots are, notably, Ishiguro [1] and Hanson [2], who have created some of the best realistic humanoid robotic heads with articulated faces. Another example is the Jules robot at Bristol labs [3] achieved in collaboration with Hanson. However, despite tremendous efforts by many researchers, these realistic robotic heads still struggle with a problem called the "Uncanny Valley" [4], [5], a phenomenon wherein people feel the appearance or behaviour of these heads is un-natural or strange. Given the effort required to build just one of these heads – careful face appearance design, mechanical design and construction efforts, and possibly hardware-dependent control algorithms – it is obviously not easy to go back and change the head upon re-evaluation. Thus, it is important to find optimum face models and behaviours before building an actual robotic face.

Computer graphics-based approaches are often used to evaluate various facial behaviours and appearance for robot heads. The three major approaches are: (1) using an LCD display itself as a robot face; (2) using a hybrid approach where an LCD display is embedded into a physical shell; (3) using a data projector to project onto a non-flat screen. Using a computer display to visualize the head is the most straightforward solution, and the display can be mounted on a robot platform to make an integrated system [6], [7]. Of course, the virtual face's physical appearance is limited by the 2D computer display.

The hybrid approach used by Bazo et al.[8] is able to display different facial features – eyes and mouth, for example – on a display embedded as part of a contoured robotic face shell, augmenting the flexibility of computer graphics with a 3D physical structure. This solution facilitates changing the face in subsequent design cycles, and is a good solution for designing "robotic-looking" as opposed to "realistic-looking" humanoid robot faces. However, the overall shape of the robot head will be limited by the shape of the 2D computer display.

The curved display approaches use abstract (cartoonish) face models: simple eyes, eyebrows, a nose and a mouth projected onto a sphere, [9], or FACS[10]-based simple face models [11], [12] projected onto an abstract version of a face mask. Similarly to Bazo et al. [8], these faces are able to convey only a small subset of the behavioral complexity available in more realistic representations, and as such can only test the capabilities of this subset. However, it is an open question as to how much and what type of information a face needs to convey in a given situation, and in some contexts simpler faces may be preferable. Conversely, for some applications a realistic face may be required. This suggests that building a platform to handle both simpler and more realistic 3D faces

such as ”Mask-bot” in Figure 1 is quite useful.

An alternative approach called Hypermask [13] which projects a face animation onto a monotone mask worn by an actor from a separated remote data projector, has the advantage of not requiring a complicated mechanism on the projection target (Hypermask itself is used for a human subject, but it could be applied to humanoid robots.) However, the projected area will be limited by the location of the remote data projector, and it is not practical for our purposes - the projector must be near the target robot face.

Additionally, another front image projection system with a dynamically modifiable soft mask face was produced by Hayashi et al.[14]. It can modify a life-size face mask to different subjects and different facial expressions, although a main drawback is that the face system requires significantly large mechanical structures behind the face. Also it has the same problems of any front video projection: specifically, it is not practical for various evaluations, especially those using Auditory-Visual (AV) speech with head motion.

In this paper, we introduce our rear-projected 3D talking head robot, “Mask-bot”. The system has the advantages of being a rear-projected system and, additionally, has the following features: it uses a realistically-shaped 3D face mask as a screen, and can project and animate a range of faces, from simple to realistic. By projecting a calibrated face animation, we produce a realistic, life-size robot head. Our work has the advantage of being able to change the robot face appearance and behavior easily, thus addressing the problem of finding optimum face models and behaviors before building hard-to-modify platforms. In addition, our system is built to explore one of our main concerns: creating robots with effective human communication skills to aid in robust, safe collaborative behaviors between robots and people.

II. System overview

The Mask-bot communication system consists of four components: the Mask-bot display system[15], the OpenHRI-based[16] speech communication system, the talking head animation system specially calibrated for the Mask-bot display, and a pan-tilt unit (PTU) control system. An overview of the system is shown in Figure 2. The system was designed especially to accommodate communication with human participants using speech as input, and talking head animation as output. Coupled with the ability to display realistic as well as simple faces, Mask-bot is able to effectively convey auditory-visual information and affective communication information.

A. Speech communication system

Our speech communication system is established by an OpenHRI platform[16]. It is equipped with English and Japanese speech recognition modules based on Julius[17], a simple keyword-to-speech response module using SEAT (Speech Event Action Transfer) script[18], an English text-to-speech (TTS) synthesis module based on the MARY TTS system[19], and a UDP trigger module which works seamlessly with the OpenHRI platform[16].

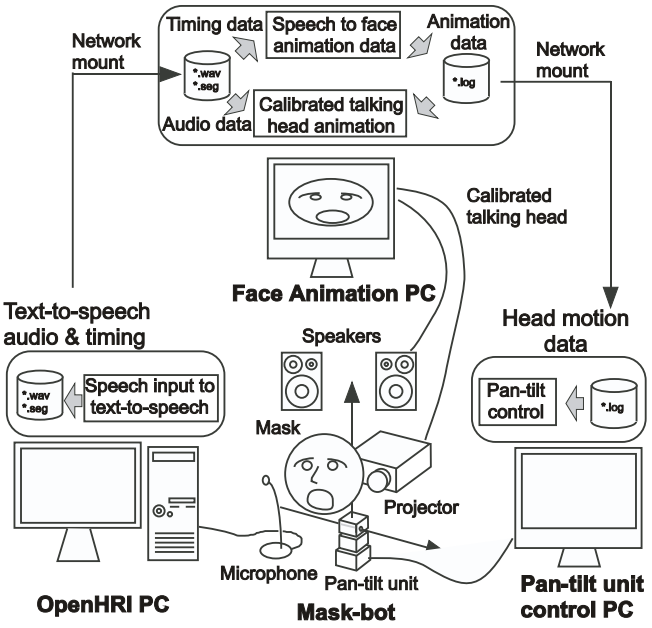

Fig. 2. *“Mask-bot” communication system diagram*

Furthermore, Japanese TTS functionality was established by a simple Japanese Kana to English phoneme conversion. Using a SEAT script written in XML format, we provided simple rules connecting some English / Japanese input keywords to specific TTS output events for greetings and sample sentences.

B. Mask-bot display system

The Mask-bot display hardware consists of three main components: 1) a monotone mask; 2) a projector; and 3) a motor-controlled base, as shown in Figure 1[15]. To make the current system as small as possible, we applied rear-projection from a projector with a fish-eye lens to the mask. A similar strategy is employed in “LightHead” from Delaunay and colleagues [11], [12] and the “curved screen face” from Hashimoto and colleagues [9]. We decided to use a portable LED projector with 200 ANSI lumens with contrast 2000:1 (K11, Acer Inc.), to use under normal indoor illumination conditions, since the smaller pocket or pico projectors (15-50 ANSI lumens) available on the market are too dark in the same conditions. We selected a mask with an embedded facial structure rather than a simplified face mask or a curved surface since our first target was a realistic life-sized robot head.

The system configuration shown in Figure 1 was mainly determined by the following constraints: the fixed projection angle and direction, the divergence properties of the fish-eye lens (x0.25), and the projector landscape (versus portrait) mode.

1) Mask screen: In our preliminary design steps, we tested various materials as potential screens with a brighter data projector (4000 ANSI lumens): e.g. thin white plastic masks, thin papers or cloths, and various semi-transparent plastics. However, most of the materials showed poor illumination performance even with the stronger projector, with the exception

of a special paint fabricated for rear-projection screens (Rear Projection Screen Goo, Goo Systems).

Mask-bot uses the front half of a transparent dummy head sprayed with this rear-projection paint on the inside of the head. Spraying on the outside surface slightly improves the look of the head as well since the paint reduces the shininess of the transparent plastic surface. However, because any exposed painted surface is easily damaged by contact with hard objects, even by a finger nail, we opted to paint only the interior surface. As a result, a 200 ANSI lumens projector used together with this rear projection painted mask can be used under normal indoor illumination.

2) Data projector with wide lens: Data projectors are usually designed to project onto a big screen from a certain distance. Therefore, it is necessary to modify the optics to project from a shorter distance, keeping system size to a minimum. Some projectors can add an optional lens to modify the projection distance, but the number of models is limited, and such options still do not match our requirements to project the life size mask from a reasonably small distance. Therefore, we make a 37mm lens mount tightly aligned with the front of the projector lens.

3) Pan-tilt unit: The system requires powerful motors to move the brighter (and thus heavier) projector along with the mask screen and the extra supporting structures. For this reason, we use a heavy-duty pan-tilt unit with 12 lbs. (5.44kg) payload capacity, the PTU-D47 by FLIR Motion Control Systems, Inc (formerly, Directed Perception). Even though this model does not have a yaw degree of freedom, we decided to use it for quick evaluation for simple head movements. The current Mask-bot output module without the pan-tilt unit (PTU) and cable weights about 1.44kg. (The projector itself is 0.61kg.) However, optimizing the mount base structure is expected to reduce this weight significantly.

The PTU is connected to a PC via serial port and controlled by a head motion playback program which loads 6 DOFs (degrees of freedom) head motion data and uses 2 DOFs to control pan and tilt in real-time. Our major concern with this unit is its operation mode: since it was designed for precise position control and not for smooth motion control, it may not be suitable for real-time control for mimicking human head motion.

C. Talking head animation system

The talking head animation system is adapted from the text-to-AV (TTAV) synthesis system we developed at MARCS auditory laboratories[20], and is based on a statistical mapping of principal component analysis (PCA) results between 3D face motion capture data and 3D face geometry data[21]. The system consists mainly of two pieces - a speech-to-face animation data engine and a talking head animation engine. The former is notified by the OpenHRI system when a new text-to-speech (TTS) sequence is created, and synthesizes PCA-based animation data from phoneme timing data provided by the TTS. This animation data is synthesized in real-time using a phoneme-to-face animation database built from speech

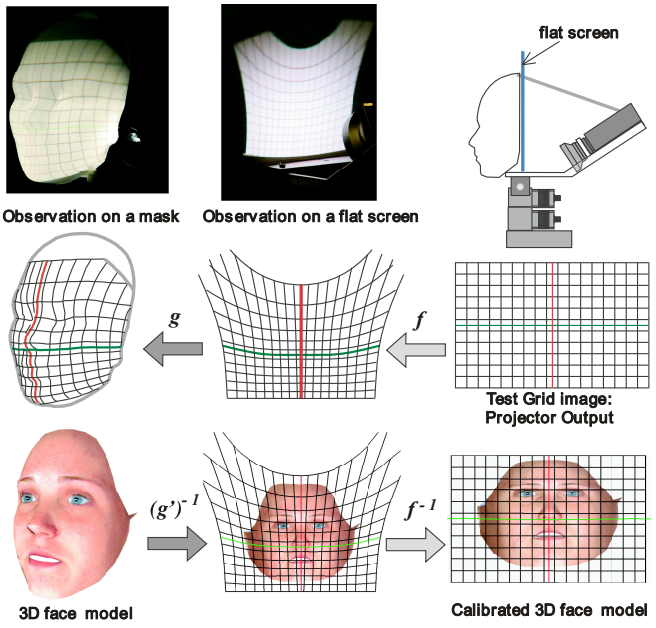

Fig. 3. *Image distortion is corrected by finding a correlation between a regular grid before and after projection onto a flat screen (f). We also correct for the distortion caused by to projecting onto a 3D face mask screen g.*

and motion capture data from an Australian English male speaker at MARCS. Finally, the animation engine plays back the talking head animation using this newly synthesized data along with synthesized speech audio from OpenHRI. Before being projected, however, the 3D model must be calibrated to account for the distortion in the system.

The animation engine can play either TTS output or recorded AV speech data, and sends a playback signal to the PTU head motion playback program to reproduce head motion. For TTS data, pre-recorded head motion is randomly selected during the phoneme-to-face animation generation process, whereas recorded AV speech data includes actual head motion.

The current TTAV system generates speech-related face motion without any emotional expressions (i.e. neutral speech), even though face models used for the animation contain various emotional expression parameters which are distributed among their principal components. However, we are planning to add emotional expressions independently from speech motion as a start, and then try to synthesize emotional speech and correlated face motion.

III. Projection Calibration

We need to account for the two main types of distortion in the current system 1) distortion from the fish-eye lens and, to a much lesser extent, the projector itself (f); 2) distortion from projecting a model intended for a 2D surface onto a 3D surface (g), and inverse distortions from 3D face model to the computer screen (bottom of Figure 3) to compensate for distortions from the projector to a 3D surface ($f + g$). We address these two cases separately.

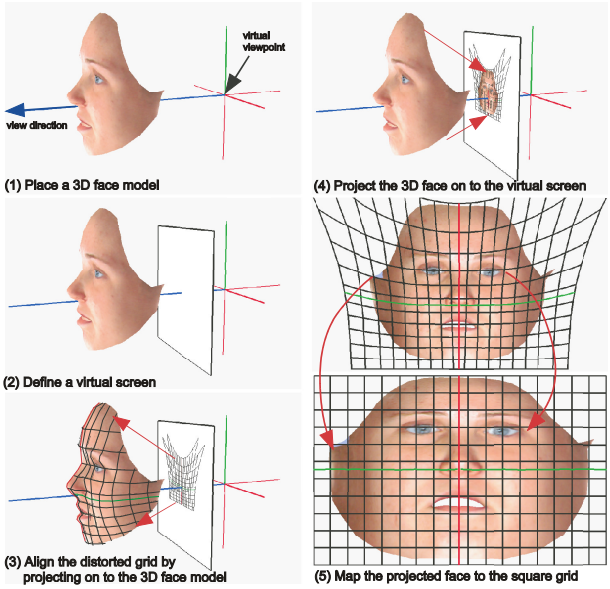

Fig. 4. *Calibration process for a 3D face model. Defining a virtual view point and a virtual screen which correspond to the projector's ideal optical center, and a flat screen used for grid distortion. (The virtual screen in this figure is located closer to the virtual view point than actual calibration procedure to visualize the process clearly.)*

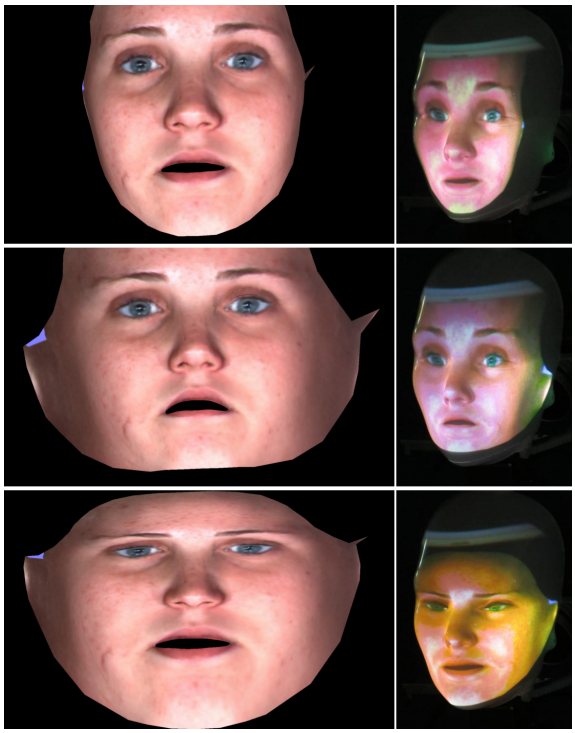

Fig. 5. *Comparison of uncalibrated images in the top two rows: top = 3D face model without calibration, middle = 3D face mask screen calibration), $(g')^{-1}$. On the bottom is the fully calibrated image. (Left column is on the computer screen and the right column is on Mask-bot)*

A. Measurement of the projected fish eye lens distortion

To separate the face mask calibration problem from the fish-eye projector problem, we put a flat screen just in front of the mask screen. We project a 2D regular grid pattern through the system to observe the resulting distortion. To do this we take a picture of a flat screen as shown in Figure 3, apply a simple Affine transformation to modify the normal aspect ratio, and then manually extract the distorted grid points. We then defined a linear mapping model f between points on the original grid and the projected grid and it's inverse mapping f^{-1} for later use.

B. 3D face mask screen calibration

Our correction for the second type of distortion – projection onto a 3D (face) screen g and its inverse mapping g^{-1}, can be obtained in a straightforward manner if we have the 3D geometry of the mask. Without this we can approximate the shape of the mask using data from the 3D face models available in our animation engine. Using 3D face geometry from one of the models, we approximate the shape of the mask to obtain $(g')^{-1}$ as shown in Figure 3. The detailed steps to obtain $(g')^{-1}$ are visualized in Figure 4 (1) to (4):

(1) First, we locate the target 3D face in virtual 3D coordinates at approximately the same configuration as the actual 3D face mask. We do this by defining an optical projection center (roughly estimated) as a virtual view point and an origin for the 3D coordinates.
(2) We define a virtual plane at the same location as a flat screen where we observed grid distortion as shown in Figure 3.
(3) We align the observed distorted grid on this virtual screen by projecting the distorted grid pattern to the 3D face from the virtual viewpoint to obtain grid output similar to that on the Mask-bot display system
(4) Then we project this 3D face mesh model onto a virtual plane. We resample all 3D points on this virtual plane to obtain the new viewpoint coordinates. The resampling is done by finding an intersecting point on this virtual screen between the view point and each 3D point.

These procedures give us a unique $(g')^{-1}$ for each 3D face model used in the animation engine.

(5) Finally, the resampled points of the 3D model from the virtual plane can be corrected for the fish eye lens and projector distortion by applying the linear map f^{-1} described earlier to obtain calibrated 3D face model.

In addition to correcting system distortions, we must accommodate the requirements of our animation engine [20]. Since the animation model uses the principal components (PCs) of the 3D face model, the original PCs are also converted to the calibrated coordinate system. Then, the animation engine can run without modification by simply loading the calibrated face model.

C. Calibration results

Figure 5 shows a comparison of LCD screen images (left) and output results on Mask-bot (right) for an uncalibrated 3D model (top row); a model with 3D face mask screen calibration

Fig. 6. *Sample animation frame output of Mask-bot*

$(g')^{-1}$ only (middle row); and a fully calibrated model, that is, with both mask screen calibration $(g')^{-1}$ and fish-eye lens distortion correction f^{-1} (bottom).

As you can see from Figure 5, the lens distortion f^{-1} affects the eye and forehead areas significantly, resulting in stretched faces that look surprised (top and middle row), compared to the fully calibrated results (bottom). Unfortunately, as we mentioned before, the 3D shape of the face animation model and the mask are not an exact match, so the final output face does not look exactly the same as the original face. However, the results of the calibration does result in a realistic 3D projected face. Figure 6 shows various mouth postures displayed on Mask-bot as a result of projecting the calibrated face model onto the 3D face mask.

IV. Animation tests

As a preliminary test of the Mask-bot system, we used a simple impression questionnaire during an internal demonstration where Mask-bot was shown under normal illumination conditions in a hallway of our institute. We used a small set of face animation sequences driven by an English TTS, including some Japanese greetings synthesized using English phonemes. People could talk to Mask-bot either in English or Japanese, and pre-defined answers were synthesized by OpenHRI and triggered by keywords.

Figure 7 shows the actual demonstration setup. The Mask-bot screen was mounted to match an adult height (∼160cm), and a scarf covered forehead area to hide a void area in the animation output. Control PCs for talking head animation and OpenHRI were located side-by-side, and the PTU was controlled by the same computer used for the animation. Built-in speakers were used for audio playback and an external USB microphone was placed in front of the Mask-bot screen to capture clear speech signals in the noisy demonstration environment.

The overall impression of the audience was quite positive: most people were very surprised that a simple mask could

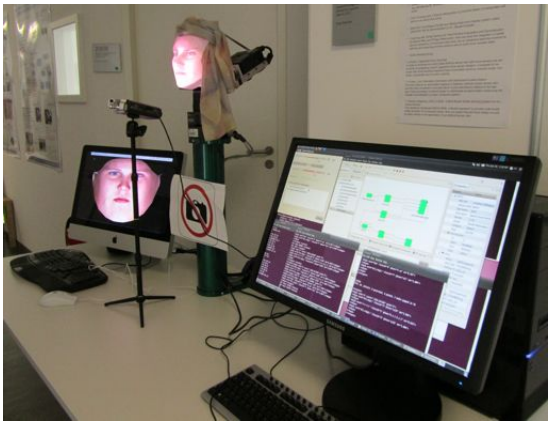

Fig. 7. *Demonstration setup of Mask-bot*

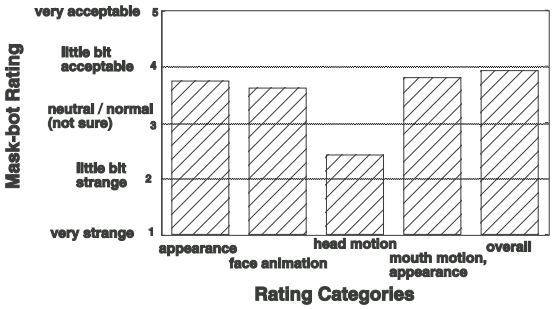

Fig. 8. *Averaged survey results for Mask-bot impressions (N=8 subjects)*

transform into a realistic face using a projected talking head, especially when seeing the calibrated face before projection on the LCD display.

Figure 8 shows subjective evaluation results of written answers from N=8 subjects. We asked subjects to rate their impressions of Mask-bot's appearance, face animation, mouth motion and mouth appearance, head motion, and an overall impression with 5 levels (1: very strange, 2: little bit strange, 3: neutral / normal (not sure), 4: little bit acceptable, and 5: very acceptable). As you can see from these averaged ratings, Mask-bot gives positive impressions except for head motion.

The pan-tilt unit was initially used to playback pre-recorded head motion sequences randomly selected for each TTS output. However, we realized that the unit was too loud, and the generated motion was not smooth enough to reproduce realistic head motion, so we disabled it most of the time. This was the major cause of the lowest impression about the head motion.

Also, some people commented that the open mouth shape looked a bit strange when observed from the mask side. Figure 9 shows such an extreme open mouth posture observed from different views. If you observe from the front, it is hard to recognize the mismatch between face animation and mask shape, but from the side view it become visible. This is a major drawback of using a realistic 3D mask. However, we will modify the mask design to improve this problem for the next generation.

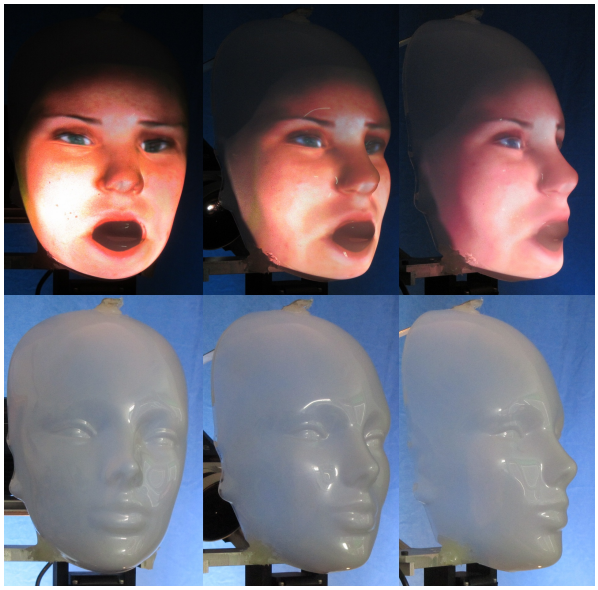

Fig. 9. *Observing mouth opening of Mask-bot from different views with rear projections (top) and mask only (bottom)*

To improve the head motion, we are modifying our control algorithm to operate with less noise, and covering the PTU with acoustic damper materials. We may also upgrade to a unit with at least pan-tilt-yaw functions and quieter motor control.

V. Conclusions

We developed a life-size talking robotic head called "Mask-bot" which is unique in its ability to project and animate a number of 3D face models, both abstract and realistic. The use of a calibrated talking head animation projected onto a realistic 3D monotone mask yields impressive three dimensional effects. We conducted an initial investigation of the public's impression of our head, and plan AV speech synthesis and human-robot communication tests. Our initial survey gave us the strong impression that the system has the potential to express richer affective facial behaviour than can be expressed by a flat computer screen. "Mask-bot" can also help identify what level of realism is necessary for implementing robust communication with humans, and in developing better face models for robots. It can also speed the development cycle for robotic heads by its ability to easily change the appearance and behavior of the model. These appearance and communication-related aspects are important issues for robots built for human-robot collaboration tasks.

Acknowledgments

This work was supported by the DFG cluster of excellence 'Cognition for Technical systems – CoTeSys' of Germany.

We acknowledge Australian Research Council (ARC) Discovery Project support (DP0666891), and ARC and National Health and Medical Research Council Special Initiatives support (TS0669874) for the support of talking head animation software.

We also acknowledge ATR-International (Kyoto, Japan) for accessing their 3D face database for supporting this research.

References

[1] H. Ishiguro, "Understanding humans by building androids," in *SIGDIAL Conference*, R. Fernández, Y. Katagiri, K. Komatani, O. Lemon, and M. Nakano, Eds. The Association for Computer Linguistics, 2010, pp. 175–175.

[2] D. Hanson, "Exploring the aesthetic range for humanoid robots," *CogSci-2006 Workshop: Toward Social Mechanisms of Android Science*, 2006.

[3] P. Jaeckel, N. Campbell, and C. Melhuish, "Facial behaviour mapping - from video footage to a robot head," *Robotics and Autonomous Systems*, vol. 56, no. 12, pp. 1042–1049, 2008.

[4] M. Mori, "The uncanny valley (in japanese)," in *Energy*, vol. 7, no. 4, 1970, pp. 33–35.

[5] F. Pollick, "In search of the uncanny valley," *http://www.psy.gla.ac.uk/∼frank (last accessed on June 15, 2011)*.

[6] R. G. Reid, R. Simmons, J. Wang, D. Busquets, C. Disalvo, K. Caffrey, S. Rosenthal, J. Mink, S. Thomas, W. Adams, T. Lauducci, M. Bugajska, D. Perzanowski, and A. Schultz, "Grace and george: Social robots at aaai," AAAI Mobile Robot Competition Workshop, Tech. Rep., 2004.

[7] C. Kroos, D. Herath, and Stelarc, "The articulated head pays attention," *Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*, pp. 357–358, 2010.

[8] D. Bazo, R. Vaidyanathan, A. Lenz, and C. Melhuish, "Design and testing of hybrid expressive face for the bert2 humanoid robot," *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 5317–5322, 2010.

[9] M. Hashimoto and D. Morooka, "Robotic facial expression using a curved surface display," *Journal of Robotics and Mechatronics*, vol. 18, no. 4, pp. 504–510, 2006.

[10] P. Ekman and W. V. Friesen, *Manual for the Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press, Inc., 1978.

[11] F. Delaunay, J. de Greeff, and T. Belpaeme, "Towards retro-projected robot faces: an alternative to mechatronic and android faces," *Robot and Human Interactive Communication (RO-MAN2009)*, pp. 306–311, 2009.

[12] F. Delaunay, J. de Greeff, and T. Belpaeme, "Lighthead robotic face," *Proceedings of the 6th International Conference on Human-robot interaction (HRI'11)*, p. 101, 2011.

[13] T. Yotsukura, S. Morishima, F. Nielsen, K. Binsted, and C. S. Pinhanez, "Hypermask - projecting a talking head onto a real object," *The Visual Computer*, vol. 18, pp. 111–120, 2002.

[14] K. Hayashi, Y. Onishi, K. Itoh, H. Miwa, and A. Takanishi, "Development and evaluation of face robot to express various face shape," in *Proceedings of the 2006 IEEE International Conference on Robotics and Automation, ICRA 2006, May 15-19, 2006, Orlando, Florida, USA*. IEEE, 2006, pp. 481–486.

[15] T. Kuratate, B. Pierce, and G. Cheng, ""mask-bot" - a life-size talking head animated robot for av speech and human-robot communication research," *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2011)*, pp. 107–112, 2011.

[16] "OpenHRI: human robot interaction middleware based on RT-component specification," *http://openhri.net (last accessed on Sep 19, 2011)*.

[17] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1691–1694, 2001.

[18] Y. Matsusaka, H. Fujii, and I. Hara, "An extensible dialogue script for robot based on unification of state transition models," *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, pp. 586–591, 2009.

[19] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003.

[20] T. Kuratate, "Text-to-av synthesis system for thinking head project," *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2008)*, pp. 191–194, 2008.

[21] T. Kuratate, E. Vatikiotis-Bateson, and H. Yehia, "Cross-subject face animation driven by facial motion mapping," *Proc. of CE2003: Advanced Design, Production and Management Systems*, pp. 971–979, 2003.