"Mask-bot"- a life-size talking head animated robot for AV speech and human-robot communication research

Takaaki Kuratate[1,2], *Brenand Pierce* ,[1] *Gordon Cheng* [1]

[1]Institute for Cognitive Systems, Technical University Munich, Germany
[2]MARCS Auditory Laboratories, University of Western Sydney, Australia
{kuratate@tum.de, t.kuratate@uws.edu.au}, bren@tum.de, gordon@tum.de

Abstract

In this paper we introduce our life-size talking head robotic system, Mask-bot, developed to support human-robot communication research. The physical system consists of a semi-transparent face mask, a portable LED projector with a fish-eye conversion lens, a pan-tilt unit and a mounting base. Mask-bot uses the mask as a screen for a talking head animation engine that is broadcast from the projector mounted behind 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, affording the means for testing new face models for AV speech synthesis and perception in life-size output without building an actual robotic head.

Index Terms: talking head, face animation, AV speech synthesis, 3D face model, humanoid robot, robotic head

1. 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[2] and Hanson [3], who have created some of the best realistic humanoid robotic heads with articulated faces. Another example is the Jules robot at Bristol labs [1] achieved in collaboration with Hanson. However, despite tremendous efforts by many researchers who work on realistic physical and virtual heads, many still suffer from problems related to the "Uncanny Valley" [4, 5].

Mori put forth the uncanny valley idea as early as 1970, formulated as hypotheses regarding people's reactions to the appearance and motion of (as-yet unrealized) human-like robots. In his theory, the uncanny valley represents a place where we lose a sense of familiarity with a human-like representation. As Mori explains, in some way our expectations of authenticity no longer match our observations, and thus the familiar becomes the strange, and is rejected. This notion has broadened over the years to include reactions to both graphical and physical humanoids, and concerns those striving for both realism and likability in their synthetic human creations. To fall into the uncanny valley today is to arrive in a place where the humanoid is imperfectly realistic, evoking negative reactions from many viewers.

Given the effort required to build just one of these articulated robot 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 behaviors before building the actual robotic face.

Figure 1: *Mask-bot system overview*

Computer graphics-based approaches are often used to evaluate various facial behaviors and appearances 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 [7, 8]. Of course, the virtual face's physical appearance is limited by the 2D computer display.

The hybrid approach used by Bazo et al.[9] 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, this approach limits the overall shape of the robot head by the shape of the 2D computer display.

The curved display approaches have early roots in classic film-based techniques such as the "singing bust" found at Disney's Haunted Mansion which projects actors' film footage onto a plaster bust. The "talking head projection" by M.I.T.[6] is most likely the first projected talking head from the research community, and extends the singing bust idea by using an actor's head movements to drive a pan-tilt unit, while also projecting recorded audio and film images onto a matching face-shaped screen.

Some more recent curved screen approaches use abstract (cartoonish) face models: In [10] Hashimoto and Morooka project a model comprised of simple eyes, eyebrows, a nose and a mouth onto a sphere, while Delaunay et al. project FACS[11]-based simple face models [12, 13] onto an abstract version of a face mask. (FACS, or the Facial Action Coding System, is a widely-used anatomically-based coding system for describing facial appearance changes.)

Similarly to Bazo et al., 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 not only be more desirable, but may be necessary. This suggests that building a platform that can easily incorporate both simpler and more realistic 3D faces such as Mask-bot is quite useful.

An alternative approach called Hypermask [14] 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 is limited by the location of the remote data projector, making it impractical for our purposes, where the projector must be near the target robot face.

Additionally, another front image projection system produced by Hayashi et al.[15] has a dynamically modifiable life-size soft face mask that can change to suit different subjects and different facial expressions. One main drawback is that the 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 AV speech with head motion.

In this paper, we introduce our rear-projected 3D talking head mask system, Mask-bot. The system has the advantages of rear-projected systems combined with these additional 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. In this paper we discuss an example using a realistic face. By projecting a calibrated face animation, we produce a realistic, life-size robot talking head. We aim to synthesize faces with higher degrees of realism along with simpler models, and to evaluate this wide range of faces for auditory-visual speech perception and facial expressions.

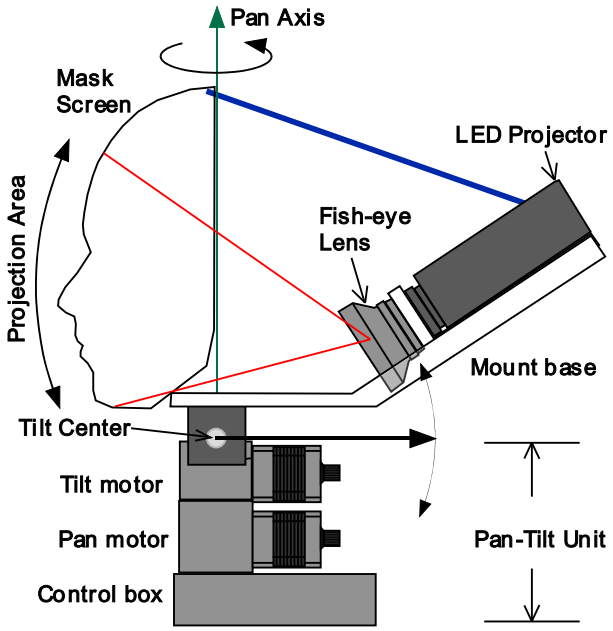

Figure 2: *Structure of Mask-bot.*

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.

2. System configuration

Our 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 2. 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 [12, 13], and the "curved screen face" from Hashimoto and colleagues.[10]. We decided to use a portable LED projector with 200 ANSI lumens with contrast 2000:1 (K11, Acer Inc.) suitable for 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 2 is mainly determined by the following constraints: the fixed projection angle and direction (usually, desktop projectors project upwards from their optical center), the divergence properties of the fish-eye lens (x0.25), and the projector landscape (versus portrait) mode.

With this configuration, roughly 85% of the computer screen can be efficiently used for animation output; 5% will be truncated by the fish-eye lens; and 10% is not used (i.e., a small unnecessary portion of the background of the talking head animation is projected, and the rest falls outside of the mask screen).

2.1. 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. 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).

The Mask-bot uses the front half of a transparent dummy head sprayed with this rear-projection paint on the inside of the head. Spraying outside of the head 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 results, a 200 ANSI lumens projector with this rear projection painted mask can be used under normal indoor illumination.

2.2. 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 requirement of projecting the life size mask from a reasonably small distance. Therefore, after testing various wide lenses, we chose a wide-angle lens, specifically a fish-eye lens (x0.25), and secured it onto our 37mm lens mount tightly aligned with the front of the projector lens.

2.3. Pan-title 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 a 5.44kg (12 lbs) 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 with simple head movements. (The current Mask-bot system excluding the pan-tilt unit and cable weights 1.44kg. The projector itself is 0.61kg. By optimizing the mount base structure, we expected to reduce this weight significantly.)

The pan-tilt unit is controlled by a host PC via an RS-232c plus USB-serial converter. Our current control program sends position information received from a talking head animation program frame-by-frame in real-time without velocity or acceleration control. As position control alone results in uneven, bumpy movements, we are adding velocity and acceleration control to produce smoother movements. These smoother movements, besides being more faithful to the original head motion, are desirable in that they reduce the operation noise of the unit's motor.

3. 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;

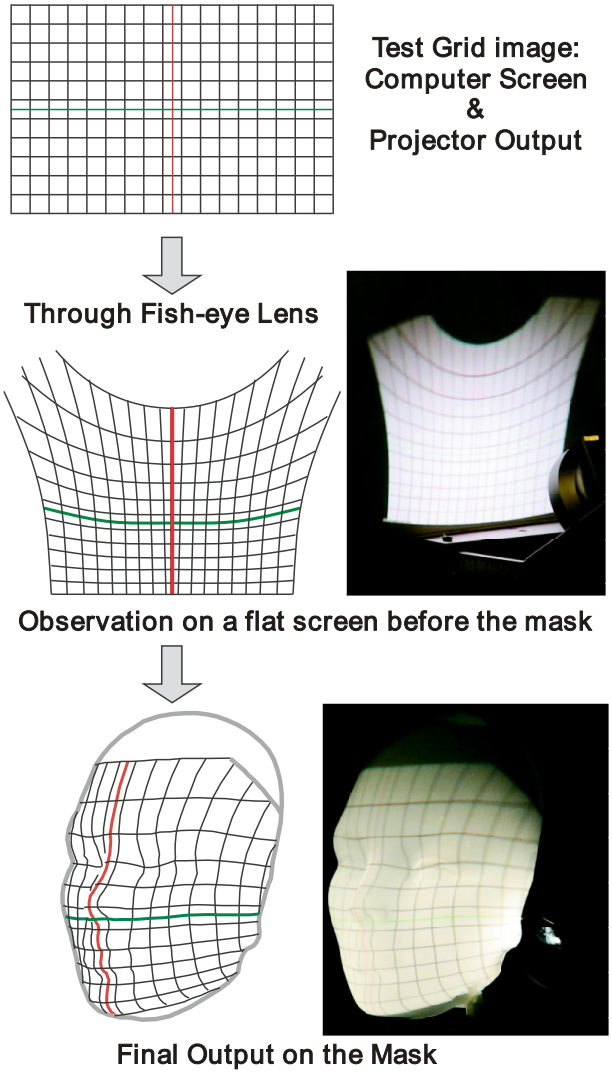

Figure 3: *(Top) Image distortion is corrected by finding a correlation between a regular grid before and after projection. We also corrects for the distortion caused by projecting onto a 3D face mask screen.*

2. Distortion from projecting a model intended for a 2D surface onto a 3D surface.

In our current system, we address these two cases separately as follows.

3.1. 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 between points on the original grid and the projected grid for later use.

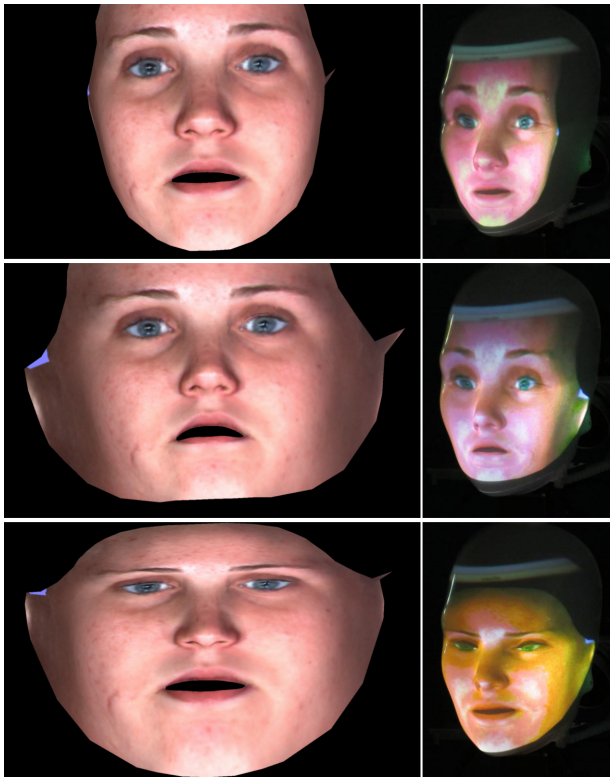

Figure 4: *Comparison of uncalibrated and calibrated images. Uncalibrated (top two rows): top = 3D face model without calibration, middle = 3D face model with mask screen calibration. Calibrated image (bottom row): with screen and lens calibration. On the computer screen (left) and projected on Mask-bot (right)*

3.2. 3D face mask screen calibration

Our correction for the second type of distortion – projection onto a 3D (face) screen – proceeds as follows. We use a 3D face model to approximate the shape of the mask. (In the future this approximate measurement could be replaced by using a high-precision 3D measurement device.) Then we project this 3D face onto a virtual plane having the same viewpoint as in the Mask-bot system. We resample all 3D points on this virtual plane to obtain new viewpoint coordinates. The resampling is done by finding a cross point on this virtual screen between a view point and each 3D point.

Now 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 described earlier.

We must also pay attention to the specifics of our animation engine [16]. We modify a 3D face model from Cartesian coordinates to a calibrated model as described above. 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.

3.3. Calibration results

Figure 4 shows a comparison between screen images on an LCD screen (left) and output results on Mask-bot (right) for an uncalibrated 3D model (top row); a model where only 3D face mask screen calibration was applied (middle row); and a model fully calibrated for both fish-eye lens distortion and mask screen distortion (bottom).

Figure 5: *Sample animation output of Mask-bot*

As you can see from Figure 4, the distortion from the fish-eye lens heavily affects the eye and forehead areas. The output face will be stretched and look "surprised" without proper lens calibration (top and middle row), while the fully calibrated image looks normal (bottom). Unfortunately, as we mentioned before, the 3D shape of the face animation model and the 3D face mask screen are not the same, so the final output does not look exactly like the original model. Despite this, the output face looks realistic in 3D.

Figure 5 shows various mouth postures displayed on Mask-bot as a result of projecting the calibrated face model onto the 3D face mask.

The animation engine can play either TTS output [16] or recorded AV speech data, and includes the ability to reproduce head motion from recorded AV data. When the head motion

data exists, it controls the Mask-bot's pan-tilt base. As is, the system configuration requires only pre-calibration of a 3D face model to use the animation system without modification. To enhance our system, it is also possible to modify the final OpenGL rendering algorithm as needed.

4. Animation tests

As a preliminary test of the Mask-bot system, we used a simple impression questionnaire during an internal Open House. For this demonstration, the Mask-bot was shown under normal illumination conditions in a hallway of our institute, and used a small set of face animation sequences driven by an English TTS, including some Japanese greetings synthesized using English phonemes.

The overall impression of the audience was quite positive: most people were very surprised that a simple mask could transform into a realistic face using a projected talking head, especially when seeing the calibrated face before projection on the LCD display. (See Figure 5, bottom left image.) However, some people realized that the open mouth shape looks a bit strange when observed from the side.

5. Discussion

The head output on the mask retains much of the realism seen on the original model on a flat screen, but has a slight "glow in the dark" effect from the mask. This effect may in fact aid in the acceptance of the head, as it takes away from the real-but-not-real-enough phenomena related to the uncanny valley that causes repulsion in viewers. In short, it is real enough to be interesting, without being too real. Also, the 3D effects of this type of 3D mask system are quite natural compared to image output from 3D TV, displays requiring 3D glasses, or view-limited 3D screens.

Results from our initial evaluation show that this system was novel to almost all participants. As such, and also perhaps partially due to the background of the participants, who were almost exclusively researchers, most people paid attention primarily to the overview of the system itself: how it was constructed, etc. We expect a different impression when detailed evaluation tests are conducted.

We are aware of the need to conduct more thorough tests on this platform, and plan to do so. It would be especially interesting to conduct AV speech evaluation tests comparing flat computer screen results with the 3D Mask-bot system. For example, will the AV speech recognition rate change with the 2D or 3D case? How will modified head motion ranging from 6 DOFs (degrees of freedom) to 2 DOFs affect the results? There are many possible ways to evaluate this system.

5.1. Technical improvements

As the current system is our initial prototype, we have already planned improvements based on our first observations. Head motion control, as previously mentioned, now uses a heavy duty pan-tilt unit: it is good for driving a large payload such as the current projector system, but suffers from two drawbacks: the motors are noisy, and the system has only 2 degrees of freedom. To present actual AV-speech-related head motion, a yaw degree of freedom is needed, and the motor needs to function quietly, similar to the system found in "TeleHead" [17].

We selected the 3D transparent face mask after testing several designs acquired from shops. It does have a few disadvantages for our purposes, however, including asymmetry and well-defined features, especially around eyes, which cause a mismatch with the talking head model. Therefore, we may explore creating masks with optimal mask properties using a 3D face database. Desirable attributes may include more generic, almost abstract face features; masks that match an average face model for gender and ethnicity; or a method to generate masks completely matched to each target face model.

An automatic procedure to calibrate the model to the mask would making it even easier to quickly change face models in Mask-bot. The same is true of calibration for lens distortion, with automatic correction making it easier to include different lenses or projectors.

LED projectors are getting brighter rapidly: we can envision in the near future using a pocket (or pico) projector to project the mask as in "Light Head" [13], or use multiple pocket projectors to illuminate seamlessly within an actual life-size head enclosure.

In addition, we are planning to add microphones and camera(s) to add auditory-visual input capability to Mask-bot. These modalities will be especially interesting for testing and evaluation human-robot communication skills.

6. 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 real. 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 behavior than using a flat computer screen. It can also help identify what level of realism is necessary for implementing robust communication with humans, and in developing better face models for robots. These appearance and communication-related aspects are important issues for robots built for human-robot collaboration tasks.

7. 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.

8. References

[1] 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.

[2] 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.

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

[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] "Talking head projection, M.I.T. council for the arts annual meeting, M.I.T., 1980." *http://www.naimark.net/projects/head.html (last accessed on June 28, 2011)*.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] I. Toshima, S. Aoki, and T. Hirahara, "Sound localization using an acoustical telepresence robot: Telehead ii." *Presence*, pp. 392–404, 2008.