Proceedings of the 2012 IEEE International Workshop on Advanced Robotics and its Social Impacts, Technische Universität München, Munich, Germany, May 21 - 23, 2012

# Development of an integrated multi-modal communication robotic face

Brennand Pierce[1], Takaaki Kuratate[1], Akinobu Maejima[2], Shigeo Morishima[2], Yosuke Matsusaka[3], Marko Durkovic[4], Klaus Diepold[4] and Gordon Cheng[1]

[1] Institute for Cognitive Systems (http://www.ics.ei.tum.de/),
[4] Institute for Data Processing(http://www.ldv.ei.tum.de/), Technische Universität München,
[2] Waseda University, Japan,[3] AIST, Japan

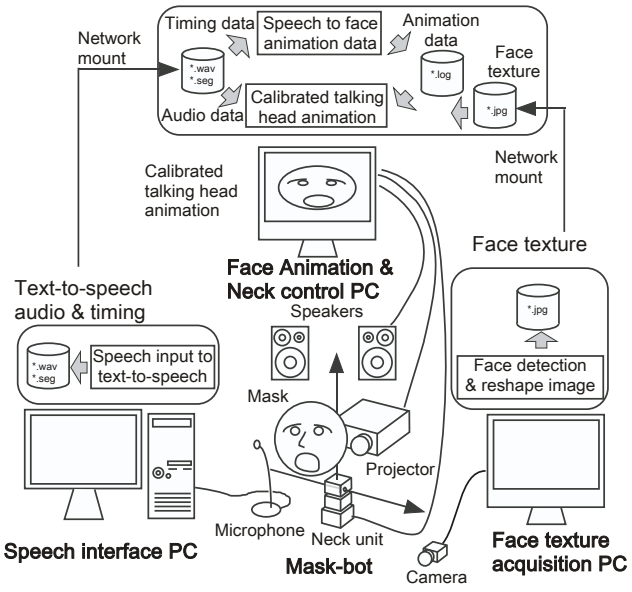


Fig. 1. Over view of the original Mask-Bot system.

***Abstract*— This paper presents an overview of the new version of our multi-model communication face "Mask-Bot", a rear-projected animated robotic head, including our display system, face animation, speech communication and sound localization.**

## I. Introduction

For robots to coexist and collaborate with humans in every day situations they need to communicate and interact in a way that is natural. For most people this means using verbal and visual communication. To study the fundamental capabilities robots need to communicate with people in the real world, we made our first communicative robotic head, Mask-Bot [1], [2].

The main difference between our head and other robotic heads is our ability to project any face onto the mask and animate the facial features, from photo-realistic people to cartoon characters. We also have a text-to-speech (TTS) system integrated into the facial animation that renders a realistic auditory-visual speech in realtime. Figure 1, gives an overview of the complete system of the original Mask-Bot.

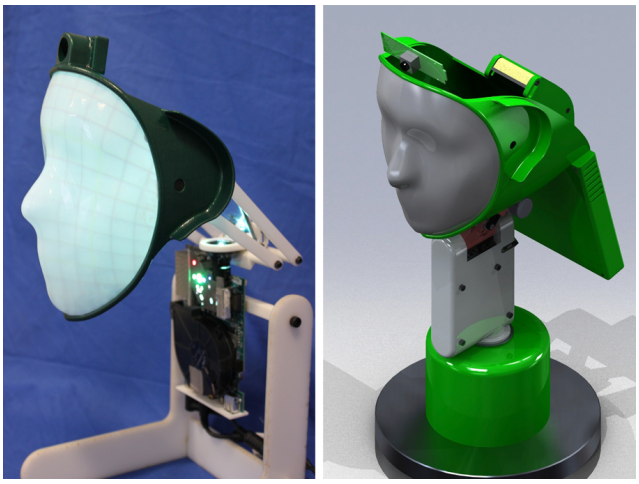

Fig. 2. Mask-Bot, Early prototype (left). Complete CAD Design (right).

## II. Mask-Bot System Design

With input from our experiments with the original Mask-Bot, we are designing a new version, shown in Figure 2. This new version will incorporate all the sub-systems into a fully integrated head, including:

- stereo speakers;
- stereo microphones;
- HD camera;
- 3 degree-of-freedom neck;
- rear projected interchangeable mask;

### A. Mask-Bot display system

The Mask-Bot display hardware consists of 4 main components: 1) a monotone mask; 2) a projector; 3) a mirror; and 4) a fish-eye lens, as shown on the left of Figure 2. To make the complete head as compact as possible, we needed to get the rear-projector as close to the mask as possible. We achieved this by first, passing the image thought a fish-eye lens, then angling the result with a mirror. A similar strategy is employed in "LightHead" from Delaunay et al.[3], [4] and the "curved screen face" from Hashimoto and colleagues [5].

For this new version of Mask-Bot we decided to use a smaller portable LED projector with 70 ANSI lumens and with contrast of 1000:1 (C112, Acer Inc.). This projector is 35% less bright then the original projector, but has the

978-1-4673-0482-5/12/$31.00 ©2012 IEEE

advantage of being 24% lighter, e.g., 138g compared to 582g. This choice was a result of evaluating the first system where we needed large motors to actuate the head, resulting in two negative consequences. The first was the overall noise of the large motors, and the second was the extra cost associated with the large pan and tilt system. Also for the current version, we made our own mask using the mean average of a large database of 3D scanned heads: we printed the mean face with a 3D printer, and then made the mask out of 1mm PETG plastic using a vacuum forming table.

### *B. Speech communication system*

Our speech communication system is established by an OpenHRI platform [6]. It is equipped with English, Japanese and German speech recognition modules based on Julius[7], a simple keyword-to-speech response module using SEAT (Speech Event Action Transfer) script[8], a TTS synthesis module based on the MARY TTS system[9], and a UDP trigger module which works seamlessly with the OpenHRI platform [6]. Using a SEAT script written in XML format, we provided simple rules connecting English/Japanese/German input keywords to specific TTS output events for greetings and sample sentences.

### *C. Face animation system*

The face animation system is adapted from the text-to-AV (TTAV) synthesis system [10], and is based on a statistical mapping of principal component analysis (PCA) results between 3D face motion capture data and 3D face geometry data [11]. 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 TTS sequence is created, and synthesizes PCAbased 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 and motion capture data from an Australian English male speaker. 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 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.

### *D. Sound Localization*

The sound localization system uses measured head-related transfer functions (HRTF) of the mask-bot. Instead of explicitly extracting localization cues like interaural time or level difference from the HRTFs and binaural recordings of the environment, the system cross-convolves the HRTF from each direction with the binaural inputs.

To estimate the positions of multiple simultaneously active sources, the localization exploits sparseness properties of sound signals and determines the HRTFs that have most likely affected the recordings. From the set of active HRTFs the localization system extracts the elevations and azimuths of all sources. The noisy localization estimates are post-processed with Sequential Monte Carlo (SMC) simulations to remove outliers and robustly track the source positions over time. For more in-depth explanation please see[12].

## III. Conclusion

In this paper we have given a general overview of our current "Mask-bot" system. This highlights that the creation of a integrated multi-modal face, requires the combination of different research areas in creating a complete system.

## Acknowledgment

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

## References

[1] T. Kuratate, B. Pierce, and G. Cheng, " Mask-bot - a life-size talking head animated robot for AV speech and human-robot communication research," *AVSP*, 2011.

[2] T. Kuratate, Y. Matsusaka, B. Pierce, and G. Cheng, "Mask-bot: A life-size robot head using talking head animation for human-robot communication," in *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, Oct. 2011, pp. 99–104. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6100842

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] Y. Matsusaka, H. Fujii, and I. Hara, "An Extensible Dialogue Script for a Robot Based on Unification of State-Transition Models," *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, vol. 2009, pp. 586–591, 2009. [Online]. Available: http://www.hindawi.com/journals/jr/2010/301923/

[9] M. S. Oder and J. Urgen, "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.

[10] 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.

[11] 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.

[12] M. Durković, T. Habigt, M. Rothbucher, and K. Diepold, "Low latency localization of multiple sound sources in reverberant environments." *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. EL392–8, Dec. 2011. [Online]. Available: http://link.aip.org/link/?JASMAN/130/EL392/1