Mary101: A Photorealistic Text-to-Audio-Visual Speech Synthesizer

Tony Ezzat

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

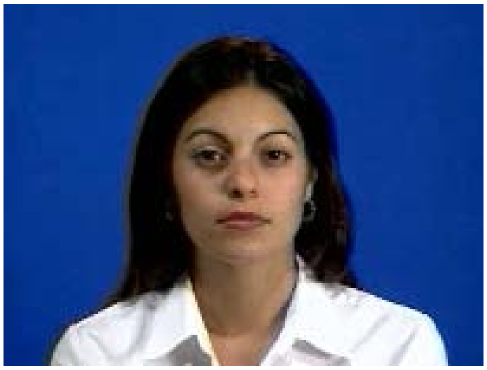

Figure 1: Mary101

The Problem: In this work, we tackle the problem of creating a photorealistic text-to-audio-visual speech synthesizer. The system should take as input any typed sentence, and produce as ouput an audio-visual movie of a photorealistic face enunciating that sentence.

Previous Work: Much of the previous work in text-to-audio-visual (TTAVS) speech synthesis [9] [2] has focused on integrating physically-based facial models with a particular speech synthesis system in order to give the impression of a ”talking face”. Some TTAVS systems have also resorted to Cyberware scanning techniques to overlay realistic-looking skin texture on top of the underlying graphics model [8].

In previous work [5], we explored an *image-based, morphing* approach to facial synthesis, in an attempt to bypass the need for any underlying 3D physical models. Our talking facial model was comprised of a collection of viseme imagery and the set of optical flow vectors [6] defining the morph transition paths [1] from every viseme to every other viseme. A many-to-one map was assumed between the set of phonemes and the set of visemes.

Proposed Work: While our earlier work made strong strides towards photorealism, it did not address the *dynamic* aspects of mouth motion. Modelling dynamics requires addressing a phenomenon termed *coarticulation* [2], in which the visual manifestation of a particular phone is affected by its preceding and following context.

In order to model lip dynamics, we propose an *unsupervised learning* framework in order to learn the parameters of a dynamic speech production model. Our new approach is composed of 3 substeps:

1. *Morpheable Model of the Lips*: we first record a training corpus of a human speaker uttering various sentences naturally. Motivated by recent progress in the creation of statistical shape-appearance models of flexible objects [4] [7] [3], we build a flexible shape appearance model of the speaker's lips. Then we analyze the entire corpus using the model, yielding a low-dimensional time-series of the lip shape for the entire corpus.

2. *HMM Speech Production Model*: Next we hypothesize a dynamic speech production model based on HMMs, motivated by the recent work of [10]. Each phone is a 3-state left-to-right HMM model with Gaussian emis-

sions. Baum-Welch learning is used to learn the paramters of each phone model from the labeled corpus calculated in step 1 above.

3. *Synthesis*: Finally we propose a synthesis algorithm to generate novel visual utterances, given input text. The appropriate HMM phone models are concatenated, and smooth set of parameters generated from the HMM using a novel synthesis algorithm that calculates the most likely emission outputs from an HMM given known state sequences.

Psychophysics: In collaboration with Gadi Geiger, We propose to perform a series of psychophysical experiments to test the output generated from our system. Specifically, we propose to perform:

1. *a Similarity Test:* This experiment will test the realism of the visual output generated by our system: Subjects will be asked whether 2 images in sequence are identical or not. The sequences presented may be 2 real sequences, 2 synthetic ones, or 1 real / 1synthetic.

2. *a Turing Test:* This experiment will also test the realism of the visual output generated by our system: In this experiment, subjects will be asked view real and synthetic visual sequences (no audio), and asked to identify explicitly which ones are synthetic and which ones are real.

3. *an Intelligibility Test:* This experiment will compare the intelligibility of the synthetic output compared to real sequences. Normal-hearing subjects will be asked to lip-read short sequences, and confusion matrices will compare the lipreading performance on the real sequences vs. the synthetic ones.

Impact: Our image-based method for facial synthesis could find potential uses in building photorealistic talking digital actors, user interface agents, and avatars. The system may also have potenial uses in very low bandwidth videoconferencing. Finally, the system may be of interest to psychologists who wish to study visual and acoustic speech production and perception.

Research Support: Research at CBCL is sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032 Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References:

[1] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *SIGGRAPH '92 Proceedings*, pages 35–42, Chicago, IL, 1992.

[2] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In N. M. Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, Tokyo, 1993.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proceedings of the European Conference on Computer Vision*, Freiburg, Germany, 1998.

[4] Tony Ezzat and Tomaso Poggio. Facial analysis and synthesis using image-based models. In *Proceedings Int. Conf. on Face and Gesture Recognition*, pages 116–121, Killington, Vermont, 1996.

[5] Tony Ezzat and Tomaso Poggio. Miketalk: A talking facial display based on morphing visemes. In *Proc. Computer Animation Conference*, pages 96–102, Philadelpha, Pennsylvania, 1998.

[6] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[7] M. Jones and T. Poggio. Multidimensional morphable models: A framework for representing and maching object classes. In *Proceedings of the International Conference on Computer Vision*, Bombay, India, 1998.

[8] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH '95 Proceedings*, pages 55–62, Los Angeles, California, August 1995.

[9] B. LeGoff and C. Benoit. A text-to-audiovisual-speech synthesizer for french. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, October 1996.

[10] T.Masuko, T.Kobayashi, M.Tamura, J.Masubuchi, and K.Tokuda. Text-to-visual speech synthesis based on parameter generation from hmm. In *Proc. ICASSP*, 1998.