Control of Speech-Related Facial Movements of an Avatar from Video

Guillaume Gibert and Catherine J. Stevens

MARCS Auditory Laboratories, University of Western Sydney,
Locked Bag 1797, Penrith NSW 2751, Australia
{g.gibert,kj.stevens}@uws.edu.au

1 Introduction

Several puppetry techniques have been recently proposed to transfer emotional facial expressions to an avatar from a user's video stream. Correspondence functions between landmarks extracted from tracking and MPEG-4 Facial Animation Parameters driving the 3D avatar's facial expressions [1] have been proposed. More recently, Saragih and colleagues [2] proposed a real-time puppetry method using only a single image of the avatar and user.

While facial expression generation may not be sensitive to small tracking errors, generation of *speech*-related facial movement would be severely impaired leading to auditory-visual integration issues. Indeed, speech is in essence a multimodal phenomenon. Compelling examples of multimodal integration are the McGurk effects [3] which are automatic perceptual phenomena appearing under incoherent multimodal information. Inaccurate transfer of facial motion can modify the sounds perceived.

The present paper describes a new method to mimic directly the user's speech facial movements from a video or a webcam.

2 Training Phase

An Australian English speaker uttered 3 times a series of non-words with a Vowel-Consonant-Vowel structure. The initial and final vowels of the non-words were identical and chosen between /a/, /i/ and /u/ (extreme lips movements) and the consonants were selected from Australian English consonants. This dataset provided the basis for building a complete articulatory model for speech production. A video was recorded consisting of a front view of a human speaker uttering the non-words against a white background. One subset of images was manually segmented i.e. the position of 68 landmarks were selected by hand for each image. An Active Shape Model (ASM) using the toolbox STASM [4] was trained on this set of images. An articulatory model was built using the method proposed by [5]. The contribution of the speech articulators (lips and jaw) was iteratively subtracted. The procedure extracted 5 articulatory parameters: one for the jaw, 3 for the lips and one for the eyebrow. Cropped images around the inner mouth area were created from the landmark positions. The DCT coefficients were computed for the red component of each image. A manual transcription was conducted between these selected images and

H. Högni Vilhjálmsson et al. (Eds.): IVA 2011, LNAI 6895, pp. 445–446, 2011.

the 5 articulatory parameters driving the tongue: jaw height, tongue body, tongue dorsum, tongue tip vertical and tongue tip horizontal. The least squares solution was then determined for the linear correspondence between the DCT coefficients and the articulatory parameter values.

3 Video Puppetry

The video puppetry animation consisted of several steps: an image was grabbed from the video stream and then cropped around the face using a face detector (Viola-Jones algorithm), then the ASM searched the best landmark positions for this image and the jaw and lip articulatory parameters were determined; finally the tongue articulatory parameters were then estimated from the DCT coefficients of the cropped images around the oral cavity area.

Acknowledgements. We thank James Heathers for manually segmenting the images. This work was supported by the Thinking Head project, a *Special Initiative* scheme of the Australian Research Council and the National Health and Medical Research Council (TS0669874).

References

1. Baptista Queiroz, R., Braun, A., Moreira, J., Cohen, M., Musse, S.R., Thielo, M.R., Samadani, R.: Reflecting User Faces in Avatars. In: Allbeck, J., et al. (eds.) IVA 2010. LNCS, vol. 6356, pp. 420–426. Springer, Heidelberg (2010)
2. Saragih, J.M., Lucey, S., Cohn, J.F.: Real-time avatar animation from a single image. Automatic Face and Gesture Recognition. Santa Barbara, CA (2011)
3. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264, 746–748 (1976)
4. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model. In: European Conference on Computer Vision, Marseille, France, pp. 504–513 (2008)
5. Reveret, L., Bailly, G., Badin, P.: MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In: 6th Int. Conference of Spoken Language Processing, ICSLP 2000, Beijing, China (2000)