Animating virtual speakers or singers from audio: lip-synching facial animation

Sascha Fagel, Gérard Bailly, Barry-John Theobald

Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 826091, 2 pages
doi:10.1155/2009/826091

Editorial

Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation

Sascha Fagel,[1] Gérard Bailly,[2] and Barry-John Theobald[3]

[1] *Berlin Institute of Technology, Straße des 17. Juni 135, 10623 Berlin, Germany*
[2] *GIPSA-LAB, 46 avenue Félix Viallet, 38031 Grenoble Cédex 01, France*
[3] *University of East Anglia, Norwich NR4 7TJ, UK*

Correspondence should be addressed to Gérard Bailly, gerard.bailly@gipsa-lab.grenoble-inp.fr

Received 17 January 2010; Accepted 17 January 2010

The aim of this special issue is to provide a detailed description of state-of-the-art systems for animating faces during speech, and identify new techniques that have recently emerged from both the audiovisual speech and computer graphics research communities. This special issue is a follow-up to the first LIPS Visual Speech Synthesis Challenge held as a special session at INTERSPEECH 2008 in Brisbane, Australia. As a motivation for the present special issue, we will report on the LIPS Challenge with respect to the synthesis techniques, and more importantly the methods and results of the subjective evaluation.

Facial speech animation can be found in a wide range of applications, among them the production of films and computer games, communication aids and tools for speech therapy, educational software, and various other kinds of information systems. The demands on facial animation differ largely with the application. Two main dimensions of the quality of speech animation can be identified: aesthetical and functional aspects. Naturalness and appeal are more connected to aesthetics; whilst intelligibility and listening effort define the function. The dimensions are orthogonal: there are animation systems of high naturalness whose output cannot be distinguished from natural video whilst the intelligibility can be very low. Conversely there are systems of clearly artificial appearance that provide intelligibility comparable to that of a natural speaker.

The techniques that are applied to animate a virtual speaker or singer range from model-based to video-based animation. The former systems use a deformable model of the face, and the latter concatenate prerecorded 2D video sequences. However, current systems—all systems described in the present issue—combine elements of both techniques.

Visual speech synthesis, that is, automating the process of matching lip movements to a prerecorded speaking or singing voice or to the output of an audio speech synthesizer, comprises at least three modules: a control model that computes articulatory trajectories from the input signal, a shape model that animates the facial geometry from computed trajectories, and an appearance model for rendering the animation by varying the colors of pixels. There are numerous solutions proposed in the literature for each of these modules. Control models exploit either direct signal-to-articulation mappings, or more complex trajectory formation systems that utilize a phonetic segmentation of the acoustic signal. Shape models vary from ad hoc parametric deformations of a 2D mesh to sophisticated 3D biomechanical models. Appearance models exploit morphing of natural images, texture blending, or more sophisticated texture models.

Comparative evaluation studies that include various visual speech synthesis systems are very rare. Usually system developers use their own specific evaluation method—if any evaluation is carried out at all. Objective or subjective results depend on the language, the linguistic material, as well as speaker-specific control, shape and appearance variables involved in data-driven approaches. Results published in the literature are thus very difficult to compare. Hence, the LIPS Challenge aimed to gather system developers in pursuit of standards for evaluating talking heads and invited them to contrast their approaches within a common framework: lip-synching a facial animation system to given acoustic signals

produced by one English speaker. Exemplars of audiovisual recordings uttered by this target speaker were available before the challenge but participants did not have to make use of this resource. One issue was thus to question if data-driven models clearly benefit from a detailed reproduction of the (training) speaker's visual signature.

Despite the fact that objective methods like RMS distance between measured and predicted facial feature points or accumulated color differences of pixels can be applied to data-driven approaches, visual speech synthesis is meant to be perceived by humans. Therefore, subjective evaluation is crucial in order to assess the quality in a reasonable manner. All submissions to this special issue were required to include a subjective evaluation. In general, subjective evaluation comprises the selection of the task for the viewers, the material—that is, the text corpus to be synthesized—and the presentation mode(s). Two tasks were included within the LIPS Challenge: one to measure intelligibility and one to assess the perceived quality of the lip synchronization. For the former task subjects were asked to transcribe an utterance, and for the latter task they were asked to rate the audiovisual coherence of audible speech articulation and visible speech movements on an MOS scale. The material to be synthesized consisted of 42 semantically unpredictable sentences (SUSs). Compared to single words used, for example, in rhyme tests or logatome tests, SUSs offer the advantage that they are well formed complete sentences constructed from real words. Furthermore, the effect of context is minimized as the keywords to be identified cannot be predicted from one another. As the evaluation should focus on articulatory movements, the subjects were presented with the lower half of the face only. This avoids distractions from mouth movements by, for example, staring or blinking eyes. All synthesized videos were to be synchronized to the given auditory speech as a prerequisite. In addition to the lip-synched audiovisual sequences, subjects were presented with the (degraded) audio alone to assess any gain in intelligibility provided by the systems. Likewise the natural video was included to access the expected upper-bound on performance. Video only was not included as SUSs are virtually impossible to lip-read. In total 30 SUSs were presented for intelligibility testing (degraded to 5 dB SNR using babble noise), and 12 SUSs were presented without degradation for rating the audiovisual synchrony.

Interestingly, three systems obtained higher intelligibility scores than the original video, with the most intelligible system being an artificial 3D head—a typical model-based system. The system with the highest MOS rating with respect to audiovisual match was a typical image-based system, which adopted a concatenative approach. Both systems achieved only moderate results with respect to the other criterion (i.e., the most intelligible system was not rated as particularly coherent, and the most coherent system was not particularly intelligible).

Feedback from viewers suggested that rating the audiovisual match was a relatively easy task; whereas subjects reported difficulties transcribing the SUS. The four multisyllabic keywords produced a high load on memory capacity. Fewer or shorter keywords will be used in future challenges. Future challenges will also aim to identify advantages and disadvantages of the abovementioned constitutive modules—trajectory formation, the shape model, and the appearance model.

There is neither a single technique for visual speech synthesis that is superior to all others, nor a single evaluation criterion that covers all essential aspects of visual speech quality. Consequently, this special issue presents a variety of systems that implement various techniques and that use different evaluation methodologies. It is the intention of the editors to foster this diversity and to encourage discussion about evaluation strategies, as both are beneficial for the research field of lip-synchronous facial animation.

Acknowledgments

The guest editors express their gratitude to the authors, reviewers, and the publisher of this special issue. they also want to thank Frédéric Elisei, Christophe Savariaux, their speaker Odette for the support of building the audiovisual speech database, and their subjects for their participation in the exertive perception test.

Sascha Fagel
Gérard Bailly
Barry-John Theobald

Author(s) Name(s)

It is very important to confirm the author(s) first and last names in order to be displayed correctly on our website as well as in the indexing databases:

Author 1

Last Name: Fagel

First Name: Sascha

Author 2

Last Name: Bailly

First Name: Gérard

Author 3

Last Name: Theobald

First Name: Barry-John