Web-enabled 3D talking avatars based on WebGL and HTML5

Jonas Beskow and Kalin Stefanov

KTH Speech, Music and Hearing
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
{beskow,kalins}@kth.se

Abstract. We describe a system for plugin-free deployment of 3D talking characters on the web. The system employs the WebGL capabilites of modern web browsers in order to produce real-time animation of speech movements, in synchrony with text-to-speech synthesis, played back using HTML5 audio functionalty. The implementation is divided into a client and a server part, where the server delivers the audio waveform and the animation tracks for lip synchronisation, and the client takes care of audio playback and rendering of the avatar in the browser.

Keywords: talking avatar, webGL, html5, text-to-speech

1 Introduction

The web as a platform for intelligent virtual agents is increasing in popularity. Traditionally, many web-based agents have employed text-based interfaces and been constrained in terms of graphical expression by browser limitations (e.g. animated GIF:s or 2D graphics), or have required the use of 3:rd party browser plug-ins. Recent web standard developments have made it possible to build advanced browser-based applications and interfaces involving streaming audio and hardware-accelerated 3D graphics without resorting to plug-ins. In this paper we describe how we leverage these recent browser advances in order to bring real-time animated talking 3D avatars with high fidelity lip synchronisation to the web.

2 Talking Avatars

People are highly sensitive to lip movements during perception of speech, and incongruent visual and auditory information is known to result in reduced intelligibility or sometimes the entire percept being altered [4]. We have been developing 3D animated avatars driven from text or speech [1] or motion capture [2] that have been shown to increase audiovisual intelligiblity of speech.

In the system presented here, we use face models generated using *FaceGen Modeller* software. The standard FaceGen models have a blendshape based facial parameterization consisting of 41 shapes, corresponding to different key poses, e.g. articulatory positions. Animation is produced by dynamically assigning weights to the different blendshapes on a frame by frame basis.

3 Client-server Architecture

The implementation is divided into a server part and a client part. The *LipSpeaker* is the client part, which runs in any WebGL-enabled web browser, and is responsible for actually rendering of the avatar to the browser window. The client is written in JavaScript, and 3D rendering is done using an intermediate highlevel graphics API [3]. Certain idle animations (eye blinks etc) are produced locally in the client. Speech animation, on the other hand, is produced server side by the *LipService* (see below). Speech animation is audio-timed in order to ensure AV synchronisation. For this to work, it is essential that accurate audio playback timing information is available. We use HTML5 audio functionality, which in our experiments has provided suficcient accuracy.

The *LipService* is the server part of the implementation. Its first function is to act as the middleware between the client and a text-to-speech service. In our current implementation we use CereProc CereVoice Cloud TTS, but in principle any TTS system that provides metadata regarding phoneme and timing information would work. LipService retrieves the speech data from the TTS server and returns an audio URL to the client.

The other function of *LipService* is to generate speech animation tracks for the avatar. This is done using a rule based system taking co-articulation and non-verbal facial movements into account [1] that takes phoneme and timing metadata as input and produces animation tracks as outout. These are sent back to the client in the form of a JSON object.

Total system latency, from synthesis request to start of animation playback is on the order of 1-2 seconds.

4 Notes

The WebGL based avatar was developed with support from the European Education, Audiovisual and Culture Executive Agency (EACEA) for use in the *LipRead* project. The avatar can be tried online at

`http://www.speech.kth.se/~kalins/projects/lipread/avatar.html`

References

1. S. Al Moubayed, J. Beskow, and B. Granstrm. Auditory-visual prominence: From intelligibilitty to behavior. *Journal on Multimodal User Interfaces*, 3(4):299–311, sep 2010.
2. S. Alexanderson and J. Beskow. Animated lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech & Language*, (In press), mar 2013.
3. Mr. Doob. Three.js. `https://github.com/mrdoob/three.js`, 2013. [Online; accessed 21-April-2013].
4. Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746748, 1976.