Automatic Visual Speech Animation

José Serra[1], João Freitas[2], Miguel Dias[2,3], and Verónica Orvalho[1]

[1]Instituto de Telecomunicações
[2]Microsoft Language Development Center, Tagus Park, Porto Salvo, Portugal
[3]ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa,Portugal
jserra@dcc.fc.up.pt, i-joaof@microsoft.com,
miguel.dias@microsoft.com, veronica.orvalho@dcc.fc.up.pt

Abstract. Visual speech animation, also known as lip synchronization, is the process of matching a speech audio file with the lips' movements of a synthetic character. Visual speech is a very demanding task, being either fully manual, which is very time consuming, or with automatic methods based on data analysis. Currently, there is still no automatic method that generates any sequence of visual speech, without requiring further fine tuning. This research focused on the problem of automatically achieving lip-sync and led to a system that relies on speech recognition to obtain the pronounced words, mapping them to the visual poses, thus automatically obtaining visual speech animation. Automatic visual speech animation has great impact in the entertainment industry, where it can reduce the time required to produce the animation of talking characters.

Keywords: Visual Speech, Automatic Animation, Audio Based, Facial Animation

1 Introduction

Speech is a key element and the most natural way to convey the ideas and thoughts of the personality behind any 3D talking character. However, speech communication is not only composed by sounds, but also by the respective facial poses and expressions, which decisively contribute to the believability of virtual characters. If speech animation is not done correctly, i.e. the movements do not resemble the ones of a person or if these are not synchronized with the audio, the viewers will find the animation awkward.

Speech can be discretized as a sequence of sounds, also known as phones, and silences. Each phone has an associated facial pose or viseme, which is the position and orientation of the visible part of the vocal tract articulators, composed by lips, teeth, jaw, tongue and cheeks. Potentially, all articulators can influence the production of a phone, however not all are visible, therefore different phones may have the same pose. Visemes are a key concept in visual speech animation. It is important to notice that when visual speech is generated by directly concatenating the visemes the resulting animation will be over-articulated. This results from co-articulation, which refers to the effect one phone, and its corresponding viseme, has over the ones around it. A digital artist has to take into account these two components when animating a speech event. Traditionally, an artist has to decompose the audio in its phones, reduce them taking into account the phonetics of the language in question, as not all influence visual speech, mark them in the sound wave using a manual technique with an appropriate audio processing tool, create the time-line with the visemes and finally control the influence of each in the final animation [1]. This is an extremely slow and time consuming process, taking about 25 to 30 minutes to animate about 10 to 15 words. Several techniques arose from the need to automatise this process, which can be divided into two categories: phone-to-viseme map, where the phones are obtained from speech recognition or text itself [2]; or based in a sub-phonetic map, where audio features are mapped to the visual poses[3].

In this demo, a phone-to-viseme map is used to create a system that relies on automatic speech recognition to obtain the orthographic and phonetic transcriptions for European Portuguese, thus generating the visual speech animation data that is then used in the Maya 3D modeling and animation authoring system. Further information can be seen in [4].

2 System Description

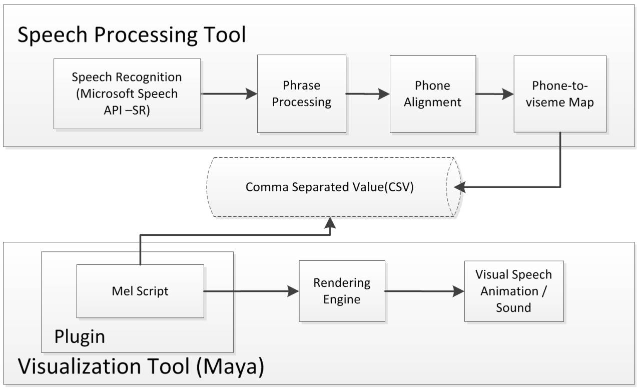
Speech Processing Tool
Speech Recognition (Microsoft Speech API –SR)
Phrase Processing
Phone Alignment
Phone-to-viseme Map
Comma Separated Value(CSV)
Mel Script
Rendering Engine
Visual Speech Animation / Sound
Plugin
Visualization Tool (Maya)

Fig. 1. Pipeline of the proposed system composed by the main modules: the speech processing tool and the plugin inserted in the visualization tool.

The proposed system is divided in two major components as shown in Fig.1: a speech processing tool, responsible for all the input analysis and choice of data that will drive the animation, and a plugin inserted in an animation engine that allowing the data flow. The pipeline of the speech processing tool works as follows: the Microsoft Speech API (SAPI 5.4) is used with a Portuguese XML grammar in command and control mode, afterwards the recognised words are passed to the phrase processing module that separates them into their phones, based on a European Portuguese phonetic transcription. With the phones and the word sound (provided by the Speech Recognition Engine) it is possible to align the phones and their times. This is done based on the phones statistical duration, where the phones duration is increased or decreased so the sum of all the word's durations is normalised to be the same as the words real audio. Finally, the phones are converted into visemes using the previously defined phone-to-viseme map for European Portuguese, thus generating all the animation data required to produce visual speech animation. The data is then stored in a comma separated value (CVS) file that is then read by a mel script in the Maya software, thus synchronizing the poses in a time line. All the poses, i.e. the visemes, weights are initially set to 1, which means the pose is always set to how it was created.

3 Results & Conclusion

To demonstrate the capabilities of the system two sentences: "olá a todos" and "animar a fala" were animated, which can be seen in the video[1]. The system is completely functional and can be used by any person, allowing a reduction of the time spent generating the visual speech by automatically synchronizing the audio and the poses. However, it is not enough to realistically produce visual speech animation, using solely the current approach, due to the need to simulate the co-articulation phenomena. Work has already been done to tackle this issue, which relies on the analysis of MFCCs audio features that vary according to the phone tone. The variations of these low level speech signal extracted features, are mapped to the viseme domain, which means the louder a person speaks the wider is the corresponding facial pose, hence higher is its weight value. In future work it is expected to effectively address the co-articulation effect towards a fully automatic system capable of generating visual speech, with appealing results.

Acknowledgments. This work is partially supported by Instituto de Telecomunicações, the projects LIFEisGAME (ref: UTA-Est/MAI/0009/2009), VERE (ref: 257695) and Golem (ref.251415, FP7-PEOPLE-2009-IAPP).

References

1. Osipa, J.: Stop Staring: Facial Modeling and Animation Done Right. John Wiley & Sons, 2010.
2. Yotsukura, T., Morishima, S., Nakamura, S.: Modelbased talking face synthesis for anthropomorphic spoken dialog agent system. In Proceedings of MULTIMEDIA '03, page 351, New York, New York, USA, November 2003. ACM Press.
3. Gutierrez-Osuna, R., Kakumanu, P., Esposito, A., Garcia, O., Bojorquez, A., Castillo, J., Rudomin, I.: Speech-driven facial animation with realistic dynamics. IEEE Transactions on Multimedia, 7(1):33-42, 2005.
4. Serra, J: Talking 3D Characters: A speech Animation and Translation Framework. *MSc Thesis*, Faculdade Ciências da Universidade do Porto, 2011

[1] http://tinyurl.com/ProporDemo