A proposal for a visual speech animation system for European Portuguese

José Serra[1,2], Manuel Ribeiro[3,4], João Freitas[3,4], Verónica Orvalho[1,2], and Miguel Sales Dias[3,4]

[1]Instituto de Telecomunicações, Porto, Portugal
[2]Department of Computer Science, Faculty of Science, University of Porto, Porto, Portugal
{jserra,veronica.orvalho}@dcc.fc.up.pt
[3]Microsoft Language Development Center, Lisbon, Portugal
{t-manrib,i-joaof,miguel.dias}@microsoft.com
[4]ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa, Portugal

Abstract. Visual speech animation, or lip synchronization, is the process of matching speech with the lip movements of a virtual character. It is a challenging task because all articulatory movements must be controlled and synchronized with the audio signal. Existing language-independent systems usually require fine tuning by an artist to avoid artefacts appearing in the animation. In this paper, we present a modular visual speech animation framework aimed at speeding up and easing the visual speech animation process as compared with traditional techniques. We demonstrate the potential of the framework by developing the first automatic visual speech automation system for European Portuguese based on the concatenation of visemes. We also present the results of a preliminary evaluation that was carried out to assess the quality of two different phoneme-to-viseme mappings devised for the language.

Keywords: visual speech animation, phoneme-to-viseme mapping, European Portuguese, virtual characters.

1 Introduction

Speech is the most natural way of conveying the ideas and thoughts of the personality of a virtual character. However, speech communication is not only composed of sounds but also of the corresponding articulatory movements and facial expressions. These poses and expressions have an important impact on the naturalness and believability of virtual characters. If speech animation is not done well, i.e. if the facial movements of the virtual character are not human-like or if the synchronization of lip movements with the audio is poor, viewers will find the animation awkward, even if they are not able to pinpoint the source of the problem.

Speech is commonly represented as a sequence of discrete sounds, or phones ('beads-on-a-string') [1]. Each phone and its abstract definition (phoneme) can be associated with a viseme, i.e. the position and orientation of the visible part of the

vocal tract articulators comprising the lips, teeth, jaw, tongue and cheeks. All articulators can influence the production of a given phone but not all changes are visible; therefore different phones may be associated with the same viseme. In computer animation, when manually animating speech events, digital artists have to create each viseme by hand. Later, they can concatenate the visemes according to the utterances they want to animate, using an interpolation scheme. Thus, manual speech animation is time-consuming and tedious. As a result, several automatic approaches have been proposed for synchronizing the audio with the visemes and for modeling co-articulation [2].

Visual speech animation can be divided into two main areas according to the way the speech input and the articulatory movements are mapped to each other: (i) phoneme-to-viseme mapping and (ii) sub-phonetic mapping. In the first case, the phonemes are obtained using text or audio analysis, mapped to visemes and organized in a timeline. The actual mapping between phonemes and visemes is important for the end result; if it is not good, the animation can appear exaggerated or have unexpected visual effects. However, a good mapping is not sufficient for high-quality speech animation, and techniques relying on diphones and triphones [3, 4] have been proposed for solving the co-articulation problem – at the cost of larger visual speech databases. Another common technique to tackle this problem involves creating a model for simulating the co-articulation effect [2, 5]. Sub-phonetic approaches, on the other hand, try to simulate continuous co-articulated speech by automatically mapping speech (represented, for instance, as feature vectors) to articulatory movements [6, 7, 8]. Using such automatic approaches makes visual speech automation faster because the individual phonemes in speech do not need to be identified as the approaches rely on a regular discretization of the continuous signal. The main problem with sub-phonetic approaches is their high sensitivity to noise. Some work in the area of visual speech animation has been done for Brazilian Portuguese [9]. However, to the best of our knowledge, research in the area has not yet been published for European Portuguese (hereafter EP).

Current challenges in the field of visual speech animation include the selection of visemes, their synchronization with audio and the modeling of co-articulation. To try to tackle any of these issues, a researcher typically has to implement a visual speech animation system from scratch, which is laborious and time-consuming process [10]. Sutton et al. [11] and Berger et al. [10] introduce the first steps towards creating modular visual speech animation frameworks. Our contribution in this area involves introducing a new concept in visual speech animation, by dividing the process into several modules. We also present the first steps towards the definition of a visual speech animation system for EP. The remainder of this paper is organized as follows. In section 2, we present two different schemes of phoneme-to-viseme mappings for EP. In section 3, we introduce our proposal for a modular visual speech animation framework. In section 4, we present results of a preliminary evaluation study that analyzed user preferences of the two mappings proposed in section 2. Finally, in section 5, we draw our conclusions and discuss our lines for further research.

2 Phoneme-to-Viseme Mapping

Phonemes are the smallest units of speech that can form contrasts between utterances. For instance, in the English minimal pair "pie" and "bye" (pronounced /paɪ/ and /baɪ/, respectively), the first consonantal sounds cause the two words to have different meanings. Therefore, we can assume that they are two distinct phonemes. The same concept can be applied in the visual domain. The visual counterpart of a phoneme is the viseme, which describes the facial and oral postures during the production of a phone. Visemes are related to the production of specific phones and are influenced by their features. Some of those features are distinctive during the production of a phone, but irrelevant in the visual domain. Nasality and voicing are examples of such features [12]. Thus, phonemes usually have a "many-to-one" relationship with visemes.

In this section, we describe our approach of mapping a 35-symbol phoneme set for EP to several classes of consonantal and vocalic visemes. Following different strategies, we created two mappings with different numbers of visemes.

The first mapping (Table 1) grouped consonants into nine different viseme classes, distinguishable primarily by the place and manner or articulation. In an attempt to reduce the number of visemes, we also created a second mapping (Table 2), which was mainly based on the place, rather than the manner, of articulation. The guttural phonemes (Table 2, Class E), whose place of articulation is near the back of the mouth, were all mapped to the same viseme, since their place and manner of articulation do not produce any relevant changes in the visual domain. The first mapping attempted to group the following types of vowels together: back vowels (Table 1, Class N), close front vowels (Table 1, Class J), close central vowels (Table 1, Class K), close-mid front and open/open-mid central vowels (Table 1, Class L), and open and open-mid front vowels (Table 1, Class M). In the second mapping, we grouped close and close-mid vowels together (Table 2, Class F), while maintaining the distinction between open-mid (Table 2, Class I) and open (Table 2, Class H) vowels. For the back vowels, we grouped close and close-mid vowels together (Table 2, Class G), and kept the open vowel separate (Table 2, Class J). This resulted in a slightly different classification of vowels, although the number of vocalic viseme classes remained unchanged. In both mappings, glides were grouped with their vocalic counterparts. So, we grouped the glide /j/ with /i/ and the glide /w/ with /u/. A final viseme, appearing as a class of its own, was considered to represent a neutral stance, or silence.

The visemes themselves were created by an experienced digital artist based on the articulatory movements made when uttering a given phoneme both on its own and in the context of other phonemes.

Table 1. First phoneme-to-viseme mapping.

Viseme Class	Phonemes
A	/m/, /b/, /p/
B	/f/, /v/
C	/d/, /n/, /t/
D	/s/, /z/
E	/ʃ/, /ʒ/
F	/ɾ/
G	/l/, /ʎ/, /ɲ/
H	/g/, /k/
I	/ʁ/
J	/ĩ/, /j/, /i/
K	/ɨ/
L	/ɐ/, /ɐ̃/, /e/, /ẽ/
M	/a/, /ɛ/
N	/ɔ/, /o/, /õ/, /u/, /ũ/, /w/
S	silence/neutral

Table 2. Second phoneme-to-viseme mapping.

Viseme Class	Phonemes
A	/m/, /b/, /p/
B	/f/, /v/
C	/d/, /n/, /t/, /l/, /ɾ/, /s/, /z/
D	/ʃ/, /ʒ/
E	/g/, /k/, /ʎ/, /ɲ/, /ʁ/
F	/ɨ/, /e/, /ẽ/, /i/, /ĩ/, /j/
G	/o/, /õ/, /u/, /ũ/, /w/
H	/ɐ/, /ɐ̃/, /a/
I	/ɛ/
J	/ɔ/
S	silence/neutral

3 Framework Description

Manually animating a talking 3D character that synchronizes facial movements with an audio signal quickly becomes impractical when the length of the utterances to be animated increases. We have created a system that can automatically handle utterances of different lengths based on a new framework for visual speech animation (see Figure 1 for the data pipeline).

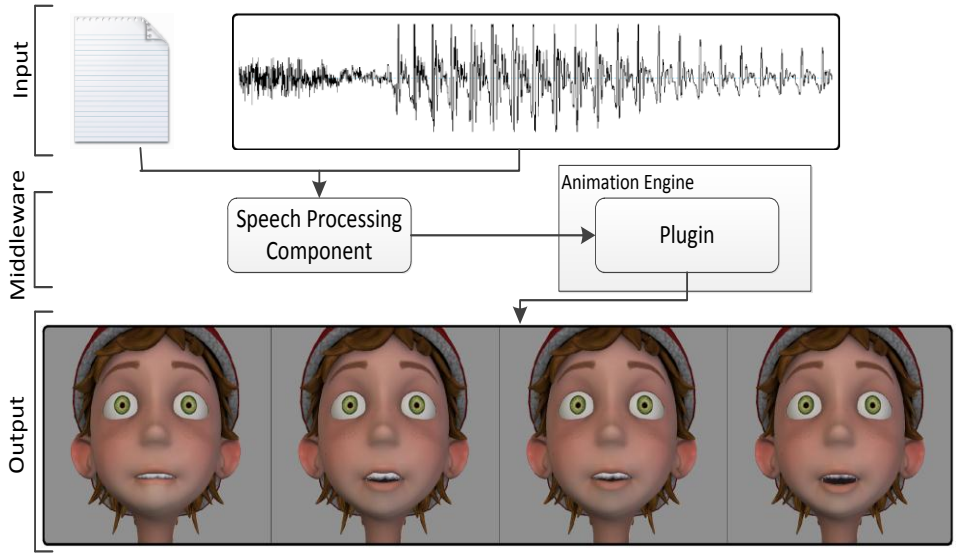

Fig. 1. Overall data pipeline of the framework. After receiving the input (audio, text or both), the speech processing tool generates the animation data and, through the plug-in, passes it on to the graphics engine that generates the animation.

The visual speech animation authoring process begins with a 3D face model in its neutral pose along with the 3D visemes, all created by a digital artist, and the utterances that we want the character to say (in the form of audio, text or both). The utterance-related information is used as direct input for animating the 3D face. The framework that makes this possible is divided into two main components: a speech processing component and a plug-in embedded in a 3D animation engine. The speech processing component analyses the utterance-related information and generates the data that drives the animation, while the plug-in acts as an external interface to the 3D animation engine. This division allows a clear distinction between the processing of the speech data and the animation of the virtual character. Thus, integrating different animation engines in our system is possible through the adaption of the plug-in. Figure 2 illustrates the modules that constitute the conceptual framework architecture. It

is important to note that the framework is independent of the system we created with it as a basis.

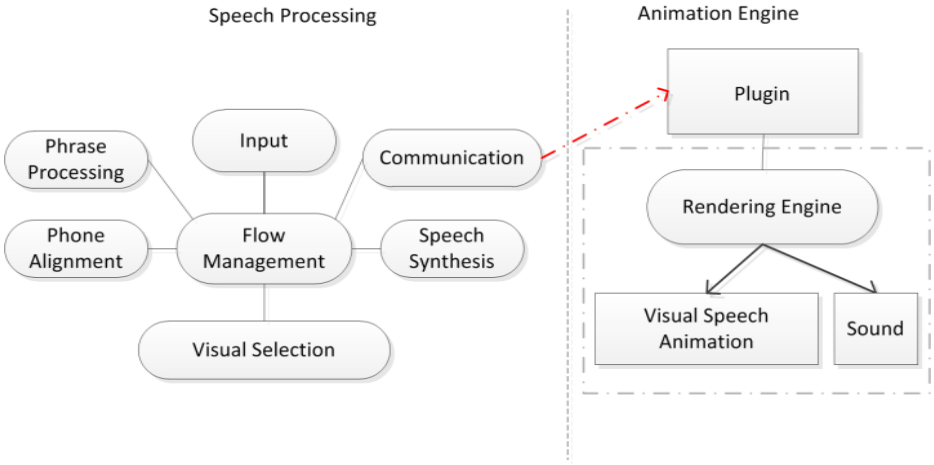

Fig. 2. Framework architecture overview. The framework is divided into the speech processing component (left) and the plug-in embedded in the animation engine (right).

3.1 Speech Processing Component

The speech processing component deals with the creation of the animation data. Central to it is the flow management module that administrates the data interaction between the different modules. The input module gets the data (audio, text or both) that will drive the animation. Phrase processing uses automatic speech recognition (ASR) to obtain the phonetic transcriptions of the input utterances. The current version of our system uses Microsoft Speech API (SAPI 5.4) [13] together with an EP phonetic lexicon developed at Microsoft. The language model of the ASR engine is essentially based on unigrams, bigrams and trigrams of common words, as well as telephone numbers, person names, business names and addresses specific to the Portuguese market. The synchronization of the audio and the visemes is handled by the phone alignment module, which guarantees that the speech is matched correctly with the lip movements. If the visualization is not correct, the animated utterances may become less understandable [14]. Techniques used in ASR and speech synthesis (TTS) are commonly used for synchronization. With ASR, for instance, a time-aligned phonetic transcription can be obtained by means of a forced alignment. It aligns a speech signal with a predefined sequence of acoustic models associated with the phonemes in question. Our current approach, on the other hand, relies on the statistical duration of phonemes; the total estimated duration of the phonemes in an utterance is normalized to be the same as the duration of the corresponding speech signal. However, in the future we intent to improve this module by changing the current approach to force align-

ment. The EP phone durations were obtained from a database of 100 hours of Portuguese speech provided by Microsoft.

The speech synthesis module is necessary when the input is text-based. To generate the audio from text input data, the current version of our system uses the EP TTS engine that comes with SAPI.

The visual selection module plays a crucial role in the framework as it is responsible for choosing the animation curves and the visemes that the virtual character will employ. There are two possible techniques that can be applied: a sub-phonetic approach or a phoneme-to-viseme mapping. In the current system, we map the phonemes directly to the corresponding visemes. The visemes were created by an artist from directly observing the mouth movements of a person speaking each phone independently. If the sub-phonetic approach is desired, only this module needs to be changed.

Finally, the communication module sends the animation data (stored in an external file) to the plug-in so that it can be displayed by the animation engine.

3.2 Animation Engine

The animation engine is divided into the plug-in and the rendering engine. The plug-in encapsulates the animation data that is sent from the communication module. The data is later translated into the final animation by the 3D rendering engine. As an animation engine, the current version of our system uses Maya, a 3D modeling and animation authoring system [15]. A cartoon character was created in Maya that relies on a bone based rig. An artist changed the default pose to create all the visemes, wich are then concatenated based on the data given by the speech processing component

4 Preliminary Evaluation

In order to analyze the impact introduced by a new phoneme-to-viseme mapping, we carried out a preliminary subjective user evaluation. The following section describes the evaluation experiment and its results.

4.1 Experiment

The evaluation was carried out using a total of 38 subjects recruited at a student fair (20 subjects) and at a multimedia systems class at the University of Porto (18 subjects) in Porto, Portugal. The subjects did not have any problems with their vision or hearing, and only 3 of them had expertise in the area of visual speech animation. They were between 11 and 69 years of age, and 79% of them were male.

The evaluation was carried out using the following three phonetically rich sentences presented to all of the subjects:

S1: A fala é um importante meio de comunicação.
‘Speech is an important means of communication.’

S2: Depois do Zé, o Ricardo joga xadrez com o Daniel.
'After Zé, Ricardo plays chess with Daniel.'
S3: O velho hoje não vê nenhum barco no mar.
'The old man does not see any ships at sea today.'

Each phonetically rich sentence was animated using the two phoneme-to-viseme mappings described in Section 3. A video with all the animations can be seen in http://youtu.be/0zZwoakx6LE. Each of the animation pairs were shown three times to the subjects, who then filled out a questionnaire according to their preferences. Corresponding to the first and second mapping, respectively, the subjects had to choose one of the following alternatives for each sentence:

L1: Strongly prefer the first animation
L2: Slightly prefer the first animation
L3: Neutral
L4: Slightly prefer the second animation
L5: Strongly prefer the second animation

4.2 Results & Discussion

Figure 3 summarizes the distribution of the preferences collected during the experiment.

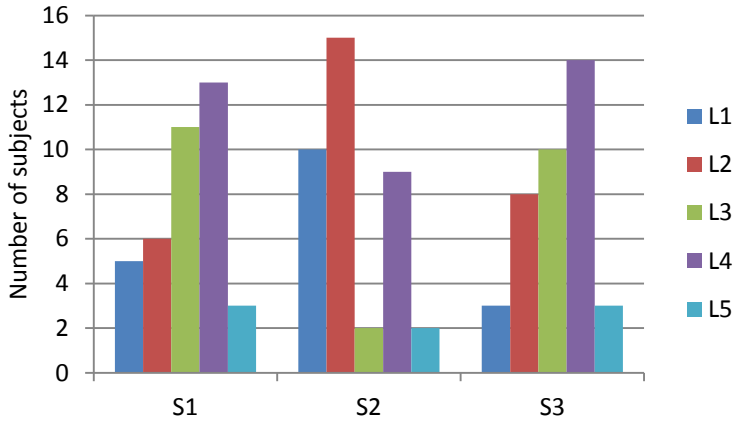

Fig. 3. The distribution of preferences for the three test sentences. L1 represents a strong preference for the first mapping and L5 a strong preference for the second mapping.

We can see from Figure 3 that the subjects slightly preferred the second mapping for S1 and S3 and that, in the case of S2, the first mapping was preferred. We can infer that the reduction in the number of viseme classes had little importance for the quality

of S1 and S3, for which most users preferred the second mapping, but that it had a negative impact on the quality of S2.

The fact that the vast majority of the preferences are centered around the more neutral alternatives (L2-L4) shows that the differences in the phoneme-to-viseme mappings affected the animations less than one might expect The differences between the two animations were rather small and, hence, our results are not fully conclusive. We can, however, conclude that the use of different mappings does influence the quality of speech animation.

5 Conclusions and Future Work

It is challenging to accurately generate a talking 3D character based on speech input or text (or both) and obtain human-like facial movements. The main contribution of this paper is the creation of the first fully automatic – albeit technologically still requiring improvements – system capable of generating visual speech for European Portuguese. The modular structure of a new visual speech animation framework makes it simple to integrate new tools into existing animation pipelines and can considerably speed up the overall visual speech animation process. Together with the framework, we also propose two different phoneme-to-viseme mappings for European Portuguese. Our preliminary evaluation experiments show that, during animation, the differences between the two mappings cause noticeable but still inconclusive changes to the quality of the animation.

In future work, we intend to improve the animation by modeling and finding a solution for the co-articulation problem, taking into account the specificity of EP. A clear starting point would be to understand the relationship between speech intensity and the visual weight distribution between visemes. As a weight model by itself is not enough to tackle the problem of co-articulation, we will also implement a co-articulation model, such as the Cohen-Massaro model [2]. We are also looking into the possibility of devising a sub-phonetic mapping method that would implicitly model co-articulation. As soon as the co-articulation problem is tackled for the case of EP, with sufficient and scientifically sound results, we will also design new objective and subjective user evaluation experiments, to validate our approach.

6 Acknowledgements

This work is partially supported by Instituto de Telecomunicações, Fundação para a Ciência e Tecnologia (SFRH/BD/79905/2011), the projects LIFEisGAME (ref: UTA-Est/MAI/0009/2009), VERE (ref: 257695), Marie Curie Golem (ref.251415, FP7-PEOPLE-2009-IAPP) and by FEDER through the Operational Program Competitiveness factors - COMPETE under the scope of QREN 5329 FalaGlobal. The authors would like to thank Xenxo Alvarez and Pedro Bastos for creating the visemes and the mel script in Maya.

References

1. Ostendorf, M., Moving beyond the 'beads-on-a-string' model of speech. In *Proceedings of IEEE ASRU-99*, Keystone, CO, USA. December 1999.
2. Cohen, M., Massaro, D., Modeling coarticulation in synthetic visual speech, Models and techniques in computer, 139-156, 1993.
3. Bregler, C., Covell, M., Slaney, M., Video Rewrite. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97, pages 353-360, New York, New York, USA, 1997.
4. Zhou, Z., Zhao, G., Pietikäinen, M., Synthesizing a talking mouth. In Proc. of the 7th Indian Conf. on Computer Vision, Graphics and Image Processing - ICVGIP '10, 211-218, New York, USA, 2010.
5. Liu, K., Ostermann, J., Optimization of an Image-Based Talking Head System. In EURASIP Journal on Audio, Speech, and Music Processing, 1-13, 2009.
6. 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.
7. Liu, J., You, M., Chen, C., Song, M.: Real-time speechdriven animation of expressive talking faces. International Journal of General Systems, 40(4):439-455, 2009.
8. Hofer, G., Yamagishi, J., Shimodaira, H., Speech-driven Lip Motion Generation with a Trajectory HMM, Proc. of Interspeech, 2314-2317, 2008.
9. Demartino, J., Pinimagalhaes, L., Violaro, F., Facial Animation based on context-dependent visemes, Computers & Graphics, Vol.30, iss.6, 2006,
10. Berger, M., Hofer, G. "Carnival - Combining Speech Technology and Computer Animation", IEEE Computer Graphics and Applications, 31, 80-89, 2011.
11. Sutton, S, Cole, R., Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, D., Cohen, M., Universal Speech Tools: The CSLY toolkit, Language, 3221-3224, 1998.
12. Erber, N., Auditory, Visual, and Auditory-visual Recognition of Consonants by Children with Normal and Impaired Hearing. Journal of Speech and Hearing Research. Vol. 15, 1972, 413-422.
13. Microsoft Speech API, http://msdn.microsoft.com/en-us/library/ee125663%28v=VS.85%29.aspx, (30 Mar 2012).
14. Verwey, J., Blake, E., The Influence of Lip Animation on the Perception of Speech in Virtual Environments, Proc. of the 8th Annual International Workshop on Presence, University College London, 163-170, 2005
15. Autodesk Maya, http://usa.autodesk.com/maya/ (30 Mar 2012).