Coarticulation and Speech Synchronization in MPEG-4 Based Facial Animation

Ricardo Leandro Parreira Duarte; Abdennour El Rhalibi; Madjid Merabti
School of Computing and Mathematical Sciences
Byrom Street, L3 3AF Liverpool
Liverpool John Moores University
R.L.Duarte@2010.ljmu.ac.uk, a.elrhalibi@ljmu.ac.uk, m.merabti@ljmu.ac.uk

DOI: http://dx.doi.org/10.1108/K-07-2014-0139

Abstract

In this paper, we present a novel coarticulation and speech synchronization framework compliant with MPEG-4 facial animation. The system we have developed uses MPEG-4 facial animation standard and other development to enable the creation, editing and playback of high resolution 3D models; MPEG-4 animation streams; and is compatible with well-known related systems such as Greta and Xface. It supports text-to-speech for dynamic speech synchronization. The framework enables real-time model simplification using quadric-based surfaces. Our coarticulation approach provides realistic and high performance lip-sync animation, based on Cohen-Massaro's model of coarticulation adapted to MPEG-4 facial animation (FA) specification. The preliminary experiments show that the coarticulation technique we have developed gives overall good and promising results when compared to related techniques.

Keywords: lip synchronization, facial animation, speech animation, coarticulation, MPEG-4.

1. Introduction

Facial animation has proved to be an immense fascination and interest by research community and the industry. As its uses is immensely broad, from the uses in the movie industry, to generate CGI characters or even to recreate real actors in CGI animation; to the games industry where coupled with several advances in both hardware and software it has permitted the creation of realistic characters that immerse the players like never before. Facial animation also reached more recently different sectors and applications, such as virtual presence and medical research. There are a few toolkits and frameworks that are dedicated to facial animation, such as Facegen , and other tools. Such broad potential and fascination around this subject has generated intense and dedicated research in the past three decades. These have led to the creation of several branches, that focus in key aspects, such as modelling (e.g. wrinkles, skin, and hair), and embodied agent systems that enhance the user's experience by incorporating virtual character with subsystems involving personality, emotions, and speech moods. These subsystems contribute to mimic our behaviours and physical appearance more realistically, and all have an important role in communication, either directly through the use of speech, or indirectly using body gestures, gaze, moods, and expressions. However, speech is the principal direct means of communication between embodied agents and the user. This justifies the efforts taken by the research community in the last thirty years toward the creation of realistic synthetic speech and lip movements for virtual characters.

In the past two decades several advances have permitted the creation of synthetic-visual audio speech for virtual characters, involving speech processing enabled with the creation of text-to-speech engines such as Mary-TTS (Schröder, 2001). With the creation of these speech engines, coarticulation also saw several advances being made (Massaro, 1993, Pelachaud, 2002, Sumedha and Nadia, 2003, Terry and Katsaggelos, 2008) which have been further complemented with facial wrinkles, intonation and matching emotional states

(Pelachaud, 2002, Benguerel and Pichora-Fuller, 1982, Sumedha and Magnenat-Thalmann, 2003).

Initial lip-sync studies attempted to concatenate phonetic segments, however it was found that phonemes do not achieve their ideal target shape at all times, due to the influence of consecutive phonetic segments on each other. Such phenomenon is known as coarticulation.

Coarticulation refers to the phonetic overlap, or changes in articulation that occurs between segments, which does not allow the segment to reach its perfect target shape. Coarticulation effects can be divided within two main phenomena, perseverative coarticulation: if the segments affected are the preceding ones; and anticipatory coarticulation if the segments are affected by the upcoming ones.

Another area which has been developed is the proposal of an MPEG-4 standard for 3D character animation [3, 7, 8]. In particular for facial animation, the MPEG-4 FA standard specification defines a set of 84 Feature Points (FP) used for both calibration and animation, of a synthetic face. Within these 84 points only 44 are used in animation, these are further sub-divided in groups (mouth, nose, cheeks, eyes, eyebrows). These feature points are controlled by 66 Facial Animation Parameters (FAP) through the application of affine transformations (translation, rotation and scaling). Additionally they can be divided into low-level and high-level parameters. The amount of displacement in each FAP is expressed in specific measurement units, called Facial Animation Parameter Units (FAPU), which represent fractions of key facial distances (Pandzic and Forchheimer, 2002, El Rhalibi et al., 2010).

MPEG-4 FA also provides a foundation to the creation of speech animation as the existent number of feature points in the mouth region permit to create, very granular and precise shapes in the lip region at all times, and provides a good framework for coarticulation research and development. Such aspect was further exploited in the projects such as; Greta (Pasquariello and Pelachaud, 2001), Balci et al. Xface (Koray, 2004), also Albrecth et al. (Albrecht et al., 2005), etc…, which not only can create realistic facial animation but can also generate visual-text-speech animation.

In this paper we introduce a novel framework compliant with MPEG-4 FA specification, capable of achieving real-time animation, FAP stream playback and recording, and is compatible with related work such as Greta (Pasquariello and Pelachaud, 2001) and Xface (Koray, 2004).

The framework is proposed to synthesize lip-sync character speech animation. It provides an easy to use MPEG-4 FA editor and player in order to create, edit, test and fine-tune facial animation and speech synchronization. In addition, it supports the development of applications for a variety of platforms, including web-based applications.

- The contributions and novelties of our work can be summarized as follows:The framework is MPEG-4 compliant and supports and extends state-of-art coarticulation approaches based on Cohen-Massaro (Massaro, 1993) model.
- The framework support FAP stream lip-sync animation as well as dynamic coarticulation based animation.
- The results achieved in our framework permit to create not only MPEG-4 FA compliant animation but also resource efficient visual-speech-animation when compared with Greta (Pasquariello and Pelachaud, 2001), Balci et al. Xface (Koray, 2004).

The remainder of the paper is organised as follows: in the next section we describe important related work in the coarticulation area. In section three we introduce the framework and our current developments - this section further sub-divides into three sub-sections in which we describe the different components of the framework: the animation approach; the XML specification which extends the FAP file specification (Pandzic and Forchheimer, 2002); our framework coarticulation pipeline; and the coarticulation model.

In section four, we evaluate and compare our work with (Massaro, 1993, King and Parent, 2005). Section five discusses the current framework features and possible future directions and further work in coarticulation in our framework. Finally section six presents our conclusions in the work undertaken, highlights areas where improvements may be achieved and where our current work provides good results.

2. Related Work

Visual speech synthesis aims at generating visual articulation movements on a talking head with accompanying speech audio. In speech theory, a viseme is defined as a basic visual unit that corresponds to a phoneme in speech. Coarticulation in its broad sense represents a condition in which an isolated viseme is influenced by, or becomes like, a preceding or following viseme. Many models have been developed to account for coarticulation in speech-driven facial animation, and to reduce the McGurk-MacDonald effect usually found within current solutions (McGurk and MacDonald, 1976). According to (Deng and Noh, 2007) we can distinguish four main speech synchronization methodologies. Data-driven and Sample-based methods, which rely solely on pre-recorded motion-data, using collections of diphonemes or triphonemes which are then concatenated (Sumedha and Nadia, 2003, Bregler et al., 1997). Learning-based methods, which, use statistical approaches such as, Hidden Markov based models (Terry and Katsaggelos, 2008) to generate visual speech.

Finally, Viseme-Driven approaches consider viseme as the basic unit data for lip-sync. We can divide further viseme-driven approaches into four sub-methods; look-ahead models (Ohman, 1967, Pelachaud, 1991), time-locked (Bell-Berti and Harris, 1979) and gestural models (Massaro, 1993, Löfqvist, 1990).

2.1 Look-ahead models

Öhman et al. (Ohman, 1967) look-ahead model, attempts to start the lips movement as soon as possible from an unprotruded vowel, towards a second onset vowel, thus time interval between the beginning of the protrusion and the end depends on the number of intervening units.

Pelachaud (Pelachaud, 1991) modified this model by assigning different ranks of deformability to each phoneme, applying to these forward and backward rules, so that a phoneme will take the lip shape of a less deformable phoneme occurring earlier or later.

2.2 Time-locked models

In time-locked models (Bell-Berti and Harris, 1979), the movement towards protrusion begins at a fixed time prior to a second vowel onset, assuming that the onset of lip-rounding is independent and not directly related to other phonetic segments besides the vowel itself. However Benguerel et al. (Benguerel and Pichora-Fuller, 1982) discovered that such independence does not apply since a vowel also influences adjacent consonants.

2.3 Gesture models

Löfqvist (Löfqvist, 1990) produced a coarticulation model that introduces the speech segment concept, where segments have dominance over the articulators, which can either increase or decrease their target value (i.e. mouth shape) over time. The target value is modulated by a dominance function to model the implicit coarticulation. Cohen and Massaro (Cohen et al., 2002, Massaro, 1998, Massaro, 1993) have further extended Löfqvist (Löfqvist, 1990) model, with the introduction of several functions. These control several aspects of the dominance functions over the articulators, such as: the dominance that a segment has over others; time offset; and duration. The integration of these parameters affects the shape of the dominance functions and consequently their target value. The control points are further complemented by an averaging function (Massaro, 1993) that determines the final target value of all segments over time.

The computational overhead of Cohen-Massaro model of coarticulation is relatively small, when compared with data-driven approaches, learning-based approaches, or even with look-ahead methods, providing excellent compromise between performance and realism.

Cohen et al. (Cohen et al., 2002), however, requires to aggregate each of the function values through the use of human subjects, machine learning algorithms and an adjustment process for each viseme in the lexicon set for a specific language (Cohen et al., 2002).

It also does not guarantee that the blending between target values starts and finishes at the same position as the silence viseme, which requires the inclusion of a new pair of silence segments between the actual speech segment set to be processed, adding a silence segment in the beginning and at the end.

The simple implementation process for Cohen-Massaro model of coarticulation coupled with the accuracy achieved have led to the development of several extensions (Albrecht et al., 2002, Cosi et al., 2003, King and Parent, 2005, Goff, 1997), or expanded the use of this model in different languages (Goff, 1997, Cosi et al., 2003).

Several authors have tried to introduce the concept of temporal resistance, in order to limit the influence of a viseme over time. Such as, Cosi et al. (Cosi et al., 2003) which integrates it within the weighted average that analyses the next and previous segment. King et al. (King and Parent, 2005) modifies the dominance function to achieve similar result, by limiting resistance to values around 2 to ensure continuity. Albrecht et al. (Albrecht et al., 2002) utilizes Kent (Kent and Minifie, 1977) findings to create temporal resistance limiting both dominance and weight average functions to limit the influence of seven segments ahead and seven segments behind, where the dominance function uses Hermite spline interpolation to restrict the support of the dominance functions.

While viseme-driven methodologies attempt to explain the different effects that occur within coarticulation, there is no known model that is capable of simulating universally coarticulation effects resulting from the many existing languages or dialects.

Viseme-driven methods require the creation of a base viseme set that enables the complete mapping of each phoneme in a language. The gathering process for each viseme set is a long complex process which has not been completely perfected, increasing the error degree of these solutions. Due to this fact, the number of visemes required to match a phonetic lexicon poses an issue, such that different authors use a different number of visemes may be matched to different phonemes. For example; Massaro et al. utilizes a set of 17 visemes (Massaro, 1998), Somansundaram (Somasundaram, 2006) utilized 12 visemes excluding silence, and Campbell et. al. (Ruth and Barbara, 1980) utilizes 14 visemes.

Throughout this section we discussed some of the issues surrounding dynamic lip-sync animation in virtual characters. Until today no solution has been found that allows us to explain completely the phenomenon of coarticulation. However we believe that viseme-driven methods are the only ones that attempt to apprehend this phenomenon using coherent mathematical models.

In the next section we will introduce our framework extension to Cohen-Massaro (Massaro, 1993) coarticulation model to create a performant and realistic coarticulation and speech synchronization in MPEG-4 based facial animation solution.

3. Charisma Framework

Our animation framework, Charisma has been created on top of the open-source game engine Ardor3D , to potentially integrate Charisma with more complex scenes or within full characters. Charisma is MPEG-4 FA compliant. Skin rendering and animation is achieved through the use of a skeletal geometrical method. This method offers good compromise between performance and realism, while being capable of creating and playback MPEG-4 based animation (Pasquariello and Pelachaud, 2001, Koray, 2004).

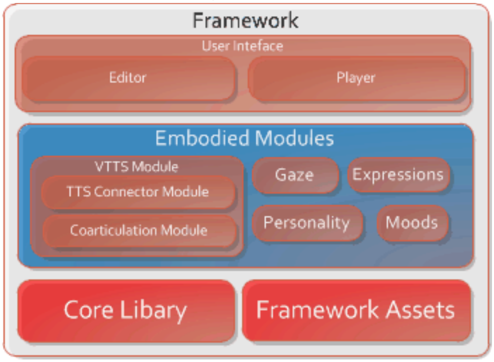

Figure 1: Our animation framework structure

Several developments have been introduced, since our previous work [9] such as layered animation manager to integrate embodied systems. The creation of an xml specification used to complement the shortcomings of FAP-stream files and to create visemes lexicon set. The creation of modular embodied layer to support our current work in coarticulation and future integration of embodied subsystems (such as gaze, expressions, emotions, etc.).

The current architecture of Charisma (Figure 1), is composed of a core Library, which contains the scene managers, animation layer manager, LoD, and mathematical aspects created specifically for the framework, while bridging through a set of interfaces with Ardor3D , but also permits the integration of other game frameworks.

The architecture is also composed of a module manager where the behaviour of a specific module is defined, and by using a listener/observer pattern it permits the integration with other modules and possible collaboration. It also permits to add new modules with ease, for further development, such as, Gaze expressions, etc. Currently the only modules integrated are the TTS module, and the coarticulation module which extends Cohen et al. (Massaro, 1993) model.

Animation is integrated using a layered approach, where each module can be animated independently, and blended with each other, with non-overlapping or overlapping feature points.

And finally the creation of new editors (Figure 2) and player applications to include our most recent work in coarticulation and speech synchronization and allow further expansion.

We will further describe throughout this section each one of these components.

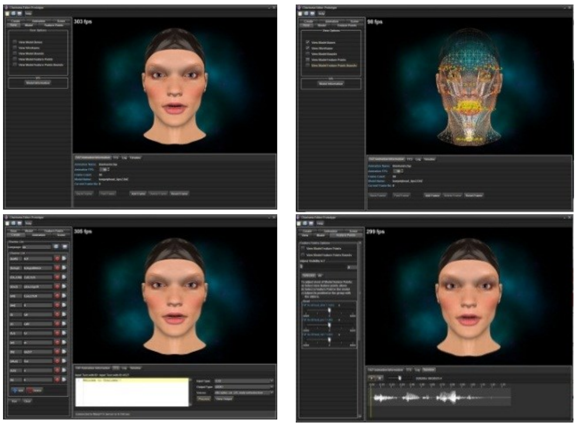

Figure 2: Our framework Editors – MPEG-4 3D model editor (Top), Viseme integration and TTS editor (Bottom).

3.1 Rendering Manager

The layered animation manager in Charisma is composed by two main sets of layers, S_{FP_g} and S_{Mod} to control the feature points.

S_{FP_g} set is composed of five layers which correspond to each feature point group, each layer controls only the feature points associated with the feature point group (FP_{gN}), such as, eyes, mouth, eyebrows, nose, cheeks, and ears. Layers in this set are animated using a direct mode approach offering compliance with MPEG-4 animation to play FAP streams.

$$S_{FP_g} = \{FP_{g1}, FP_{g2}, \dots, FP_{gN}\} \qquad (1)$$

$$S_{Mod} = \{L_{vis}, L_{exp}, L_{gaze,\dots,} L_n\} \qquad (2)$$

$$L = \{S_{Mod} \cup S_{FP_g}\} \qquad (3)$$

$$B = LERP(L) \qquad (4)$$

S_{Mod} is composed of modules that permit dynamic animation defined by specific functions specified by the model. These components are responsible for modules like coarticulation(L_{vis}), gaze (L_{gaze}), expressions (L_{exp}), etc. Animation in each of these modules is done by using a dynamic key-frame approach where states are generated dynamically by the model and integrated within the animation layer.

This approach permits the integration of FAP streams created by the animator; (such as nod, head rotation, eyes blink, etc.), whilst dynamic animation is being played, such as dynamically generated coarticulation, expressions, gaze, and other sub-systems; Blending between each of these layers is achieved using LERP/SLERP functions (4). The blend weight is defined by default so that layers in S_{Mod} set completely override layers in S_{FP_g}, but layers within S_{Mod} have a 0.5 blend weight. Weight blends can also be set by the animator.

3.2 Viseme creation and XML specification

We extended Cohen-Massaro model (Massaro, 1993) to create synthetic visual-to-speech. This approach requires the creation of a viseme set that matches phonemes for a specific language. This is achieved with our own defined XML specification, created specifically to handle FAP-stream playback with specific settings (such as; audio-file, model settings, etc.), and permits to create, modify, and visualise visemes sets for different languages and matches these to phoneme sets (Figure 2, bottom-left screenshot).

Currently, we do not have any means to capture precisely visemes shapes and associate these to phonemes, which has led us to use known visemes lists from other authors, such as Annosoft's (Annosoft) and Somasundaram's visemes set (Somasundaram, 2006) for the English language. Each set were modelled using our editor, by associating FAP values to each feature point involved and mapped to each phoneme in order to use with a text-to-speech engine.

However since each set has different mapping to phonemes and different number of visemes they will require further evaluation, using different words against a human subject.

3.3 Framework Coarticulation Pipeline

In section two we provided a broad introduction to several coarticulation methodologies and the latest relevant work within the coarticulation area.

As discussed above, the embodied layer in our framework operates using a modular approach, and includes coarticulation and speech synchronization components. Each component is handled by a class (the main processor), that is responsible for the initialization and termination of each component, and for any cooperation between components. Each component is composed by a request and a notifier class using, a listener/observer pattern that are responsible for communicating with the different parts of the framework, such as: other modules; the main processor; or the animation manager.

The main processor handles notifications and requests between each module. Visual-text-to-speech is integrated by using two different nested components, the coarticulation and TTS (text-to-speech) connector component (Figure 3).

The TTS module is responsible for the connection with a TTS server. In this work we use MaryTTS (Schröder, 2001) which is capable of creating synthetic text-to-speech for different languages, with support for different voices, and different output formats (which are useful for debugging).

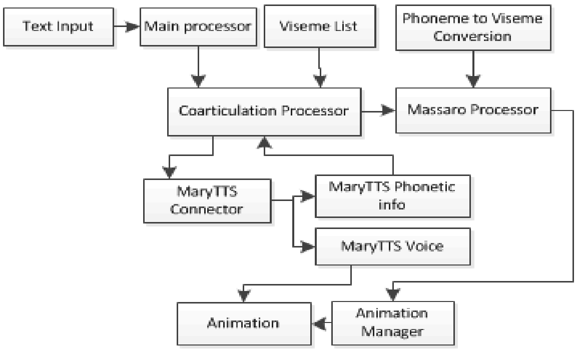

Figure 3: Our Framework Coarticulation pipeline.

The coarticulation model, which applies some of the functionality from Massaro processor, which contains the averaging function (6) and the averaging function (5) from (Massaro, 1993), initialises the TTS module and serves as a bridge between the main module processor, other modules as shown in figure 3, and the rest of the framework, by handling all the requests and notifications to other layers. To be able to create synthetic coarticulation, this module receives an input text, a selected voice, and sends two requests to the TTS module: an audio request; and a request to receive the phonemes, encapsulating their duration and prosody (i.e. the rhythm, stress, and intonation of speech), which is received in *MaryXML* format.

The coarticulation module parses the *MaryXML* file returned and matches intonation information, with the time and phoneme, to a specific viseme, creating a new set of visemes which is then used for the coarticulation.

A complete animation sequence is constructed by combining the viseme transitions provided by the coarticulation and synchronizing the visual stream with the audio stream provided by the TTS.

3.4 Coarticulation Model

The coarticulation model used in our framework extends Cohen-Massaro model (Massaro, 1993), which offers good results and efficiency. It calculates a curve representing the lip and chin movements based on the following tuple set $T = (t, \propto_{sp}, \theta, \tau, c, T_{sp})$.

These functions are integrated within a dominance function $D_{sp}(t)$ (5) and a weighted average function $F_p(t)$ (6) that gives the final target value for a viseme during its time interval.

$$D_{sp}(t) = \begin{cases} \propto_{sp} e^{-\theta \leftarrow sp|\tau|^{c}}, \text{if } \tau \geq 0 \\ \propto_{sp} e^{-\theta \rightarrow sp|\tau|^{c}}, \text{if } \tau < 0 \end{cases} \qquad (5)$$

$$\tau = t_{start\ s} + \frac{duration_s}{2} + t_{0sp} - t \qquad (6)$$

$$F_p(t) = \frac{\sum_{s=1}^{N}\left(D_{sp}(t) \times T_{sp}\right)}{\sum_{s=1}^{N} D_{sp}(t)} \qquad (7)$$

$$V_{area} = \pi \cdot V_{width} \cdot V_{height} \qquad (8)$$

In (5), $\propto_{sp}$ represents the peak magnitude of the segment pair during time *t*. θ represents the rate of magnitude up to the segment centre ($\theta \leftarrow sp$) and the rate of falloff after the peak ($\theta \rightarrow sp$). The function *c* is usually left with a constant value of 1.

In (6), function τ calculates the temporal distance between the segment centre and its t_{0sp} time offset, at time *t*.

The average weighted function $F_p(t)$ (7) was introduced by Cohen et al.(Massaro, 1993) which calculates the final target value by including all segments at time *t*.

As Cohen et al. described in (Cohen et al., 2002), $\propto_{sp}, \theta$ values for each segment are created by the use of lengthy evaluation and the use of machine learning algorithms. This is an expensive process that requires several tools, native speakers and the aggregation of all this information to create these values and associate them with a viseme set.

In our work we propose to automate all the aggregation of these parameters in order to diminish the cost associated to the creation of a talking head. We calculate the width and height of viseme in the framework active set, to apply an elliptic area calculation (8) to decide how high the peak magnitude of a viseme is.

This is further applied to (5) and consequently (7), first introduced in (Massaro, 1993).

This idea is based on the definition of peak magnitude given by Cohen et al. (Massaro, 1993), where the peak magnitude "represents the relative importance of the articulator reaching its target value for a segment". By using (8) we calculate with great precision the total area that viseme occupies, once all areas are calculated, the values are then normalised between [0..1], where visemes with greatest mouth opening have higher peak magnitude and visemes with total smallest mouth opening have smaller peak magnitude values.

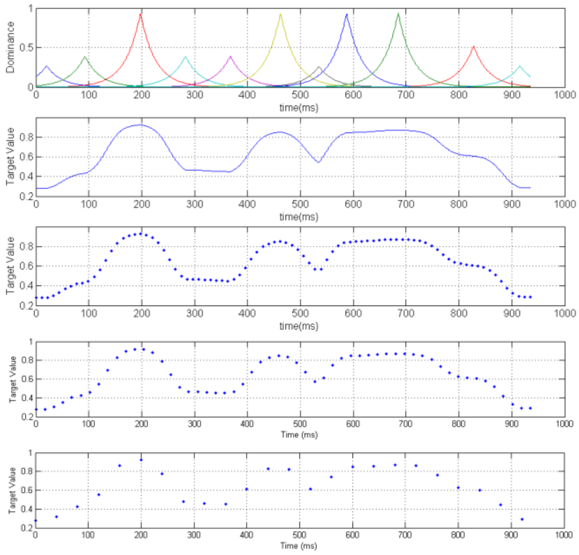

Figure 4: Dominance and target values of "Come here John" utterance, the 3 bottom graphs have been sampled to 90fps, 60fps and 25fps respectively.

In our model, the function θ is left as a constant value for now, but it will be further developed to achieve smoother dominance curves before the peak and falloff of the dominance curve.

Once $F_p(t)$ values are calculated a transform set for each feature point is created for the complete utterance.

The layered animation manager works at a fixed desired frame rate, which is set constant during the animation, while this value can be modified it is left by default at 90fps. This frame rate is then used to reduce the number of transforms necessary to create the lips movement while ensuring a smoother animation (Figure 4).

During our sample utterance "Come here John", with the duration 1066 milliseconds approximately, the original model would require 10660 transforms to synthesize all movements necessary for the 10 feature points existent in the mouth (El Rhalibi et al., 2010). Since we limit the number of transforms created to synthesise the viseme lips to match the frame rate of the animation manager, this number is reduced to 970 transforms at 90fps, 670 transforms at 60fps and 270 transforms at 25fps (Figure 4).

During our tests with several utterances of different sizes, we concluded that if the frame rate never drops below 60 fps, the animation quality is preserved and the shape of the curve for lips and chin resultant from the $F_p(t)$ values is also maintained. This is illustrated in Figure 4, where the part of the detail that is kept at 60fps disappears around 530 milliseconds creating an abrupt lip closure and raising immediately after, whilst at 60 and 90fps most of the detail is preserved in the same area, this occurs with utterances that have abrupt closures followed by lip openings.

4. Evaluation

As mentioned before our primary goals reside within the creation of a complete dynamic talking head, which can generate speech and lip movements from different visemes set with minimal effort and completely automatically.

We have created an automatic method to calculate $\propto_{sp}$ (peak magnitude) by analysing each viseme in the set, which extends Cohen-Massaro model. By using the visemes set created from (Annosoft, Somasundaram, 2006). If we analyse the results in Figure 5 we can see that we have dominance curves for all the utterances, which provide similar overall outcome as Cohen-Massaro's, and the overall shape of the $F_p(t)$ (7) is preserved. However, 's' and 't' have similar dominances as they have the same viseme shape in the used viseme set.

In Figure 6, we highlight the difference between our approach and King's (King and Parent, 2005). We can see the overall shape of $F_p(t)$ (7), in our approach, is also preserved despite the significant changes in the way dominance functions are calculated in (King and Parent, 2005) when compared with the original work in Cohen et al.(Massaro, 1993). While our approach leads to higher values of $F_p(t)$, it will however achieve similar mouth movements if both use the same viseme set, with less feature points, thus using less data to produce the same target values. While we have limited data to compare directly our work with (Albrecht et al., 2002, King and Parent, 2005, Massaro, 1993), in addition other factors can also affect the outcome of the evaluation, when we compare our results directly with related work; such as the viseme set used and the TTS engine used, since they can generate different times and for each phonetic segment. In this context,
the preliminary experiments show the coarticulation technique we have proposed gives promising results overall, in comparison to related techniques (King and Parent, 2005, Massaro, 1993, Pelachaud, 1991).

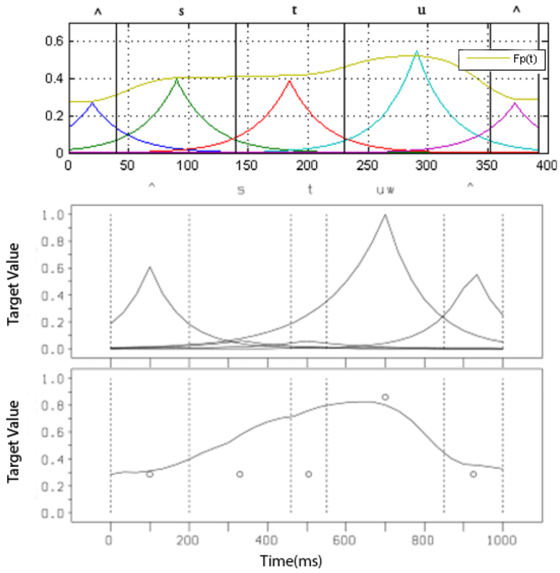

Figure 5: $D_{sp}(t)$ and Fp(t) plots for the utterance "stew", our automated system is shown in the top graph. Cohen et al. results from (Massaro, 1993) is shown in the 2 bottom graphs.

In Figure 6, we highlight the difference between our approach and King's (King and Parent, 2005). We can see the overall shape of $F_p(t)$ (7), in our approach, is also preserved despite the significant changes in the way dominance functions are calculated in (King and Parent, 2005) when compared with the original work in Cohen et al.(Massaro, 1993). While our approach leads to higher values of $F_p(t)$, it will however achieve similar mouth movements if both use the same viseme set, with less feature points, thus using less data to produce the same target values. While we have limited data to compare directly our work with (Albrecht et al., 2002, King and Parent, 2005, Massaro, 1993). In addition other factors can

also affect the outcome of the evaluation, when we compare our results directly with related work; such as the viseme set used and the TTS engine used, since they can generate different times and for each phonetic segment. In this context, the preliminary experiments show the coarticulation technique we have proposed gives promising results overall, in comparison to related techniques (King and Parent, 2005, Massaro, 1993, Pelachaud, 1991).

Our coarticulation approach allows coarticulation accuracy matching; supports automatic gathering of coarticulation parameters in particular for the dominance function; is shape preserving despite using different viseme sets; supports frame-rate controlled performance; involves less lip feature points and less data for the same coarticulation shape results; allows frame-rate adaptations whilst preserving shape; is MPEG-4 compliant (and is subsequently compatible with other MPEG-4 related work such as (Koray, 2004, Pelachaud, 2002)); and supports real-time speech synchronization and animation using a TTS engine.

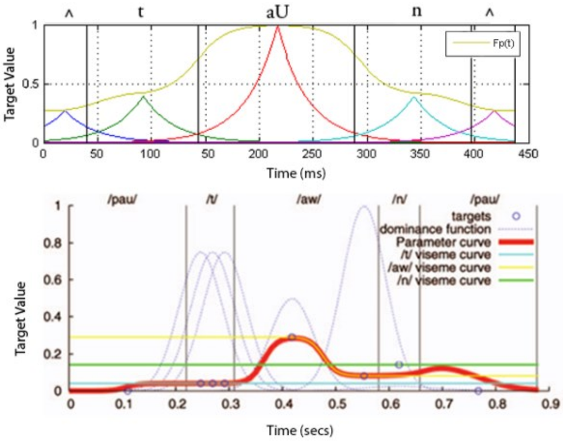

Figure 6: $D_{sp}(t)$ and Fp(t) plots for the utterance “town”, our automated system above, and King et al. results from (King and Parent, 2005).

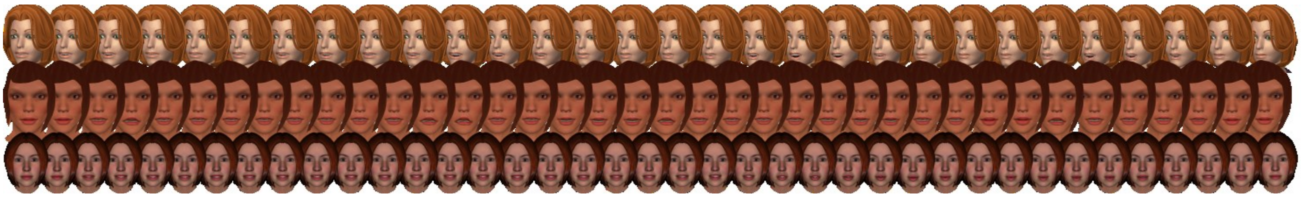

Figure 7: Comparison of our solution using coarticulation (middle row) with Greta (Pasquariello and Pelachaud, 2001) (top row) and Xface (Koray, 2004) (bottom row) animating utterance “This city”.

Our animation framework introduces a coarticulation and speech synchronization in MPEG-4 based facial animation technique. In Figure 7, we compared our results with the output created with the coarticulation system in Greta [9] and their FAP stream output played in Xface (Koray, 2004). We use the utterance “this city”, since we have some Greta FAP stream benchmarks using this utterance. We removed some head motion from the existent FAP stream, for example the head tilting, to allow the comparisons to be made only based on the utterance.

We can observe in figure 7 that both Greta and Charisma produce smooth speech animation, while Xface produces more mouth movements for the same number of visemes, which shows an indication to that their methodology is interpolating the visemes positions without considering coarticulation effects. While Greta and Charisma produce a more smooth animation it is possible to observe that Greta produces a more obvious transition between visemes, which might be due, several factors, such as better speech interpolation, or better viseme data. Since we could not obtain the source code for this version of Greta we could not attempt to convert the viseme data from Greta to Charisma.

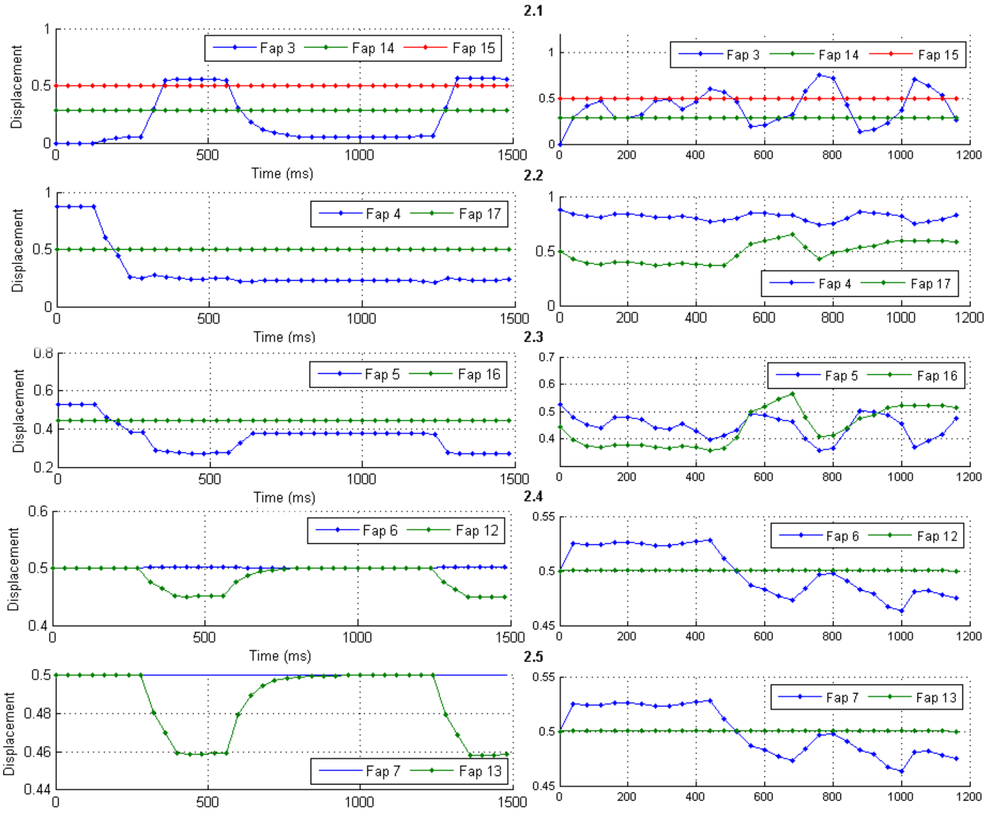

Figure 8: Lips feature points (from 2.1 to 2.5) FAPs values for Charisma (left column) vs. Greta's (right column), for utterance "This city in Szechuan".

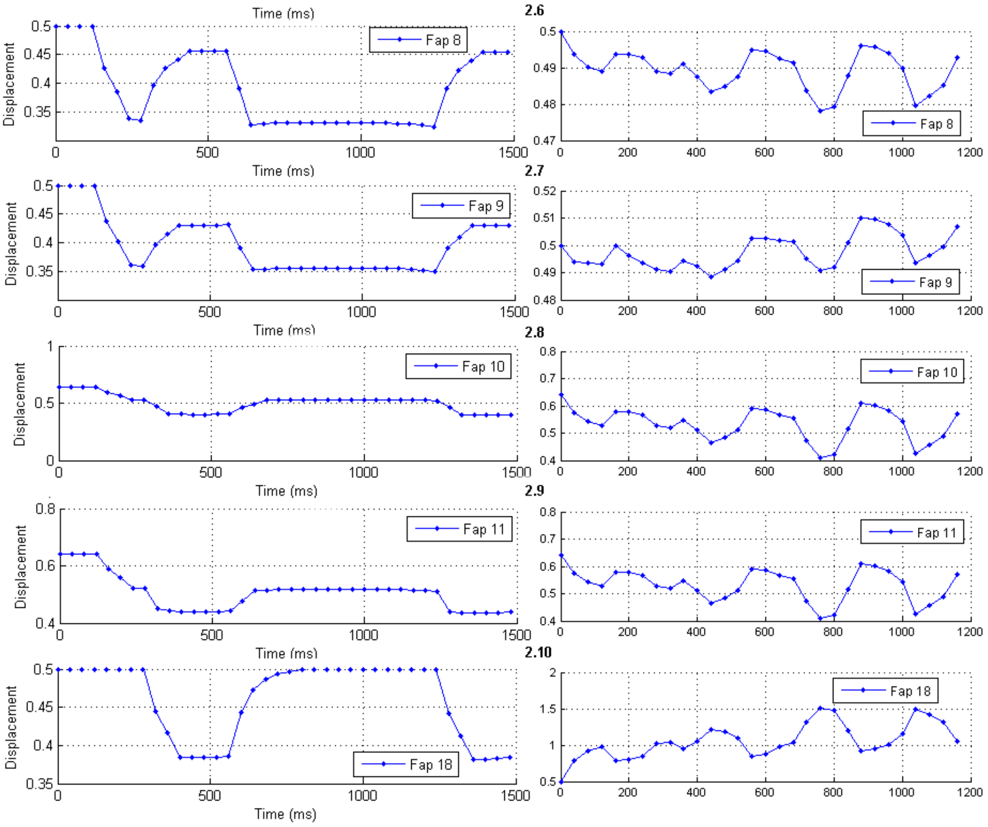

Figure 9: Lips feature points (from 2.6 to 2.10) FAPs values for Charisma (left column) vs. Greta's (right column), for utterance "*This city in Szechuan*".

In figures 8 and 9 we normalized the displacements data for the utterance "*This city in Szechuan*", and compared both Charisma and Greta. At this stage we could not compare it also with Xface since Xface was not exporting correctly coarticulation output to FAP. In these figures we compare animation smoothness between both frameworks.

By observing the plots in the feature points 2.1, 2.3, 2.4, 2.5, 2.6, 2.7, 2.9 and 2.10, we can take that both charisma and produce smooth interpolation between each frame and from one viseme to another, whilst using different coarticulation methodologies. It is possible to observe in the plots that Greta produces animation that resembles more viseme interpolation than Charisma.

In plots for feature points 2.2, 2.8, 2.10 have little resemblance or no resemblance with Charisma, this is due the usage of different visemes set, and due the way each viseme was integrated within each framework, where the usage of the outer lips was used more extensively in one framework than other.

At the current stage of our work we are creating an evaluation framework to compare how the plots for specific utterances would be represented in real subjects, and create a more objective comparison.

While, at this stage the comparison between our system and real subject is not possible, it is possible to observe that our method of aggregating peak magnitude parameters and the use of a transformer limiter permits, by automating the creation of peak magnitudes and reducing transforms calculation sizes, to obtain similar results whilst reducing greatly the integration cost involved in Cohen-Massaro model (Massaro, 1993). This result is confirmed by lip shape animation in Figure 7.

In Figure 8 and 9, we illustrate the effects of coarticulation on FAPs displacement in our framework (Figure 8 and 9), which demonstrate that our methodology can produce smooth animation, it also shows that the differences within the viseme set between each framework does not allow to produce more conclusive results.

5. Discussion

In this research, we focused on integrating the steps to create a dynamic talking head in our animation framework. We decided to implement Cohen-Massaro model (Massaro, 1993) , adapted to MPEG-4, due to its flexibility, ease of implementation, and above all the high quality of the results it achieves (Cohen et al., 2002). In section two we presented the related work (Albrecht et al., 2002, King and Parent, 2005) that extends this approach and attempts to solve some of the issues present in the original version. The improvements include shape smoothness, temporal resistance, and achieving automated gathering of some of the parameters defined in the dominance function (5). In this paper we presented our efficient approach to automate the creation of $\propto_{sp}$ in (5) which gives overall good and promising results as discussed in section 4. Our future work will focus in creating an automatic lip parameterisation model compliant with MPEG-4 and capable of further dimension reduction, possibly using principal component analysis (PCA).

We will also attempt ways to automate the rate of magnitude $(\theta \leftarrow sp)$ and the rate of falloff $(\theta \rightarrow sp)$ based preceding and following segments that exist in the utterance, and finally we intend to create our own temporal resistance function to limit $D_{sp}(t)$ in time following Kent et al. (Kent and Minifie, 1977) segment spatial influence limit of six segments to (8). We also intend to replace the necessary silence visemes that wrap each utterance to force (7) to simulate the speech starting from a rest position and stop at a rest position. We feel that changes need to be made to (7) to accommodate these issues, but above all we feel that the exponential functions found in (5) are not suited for these tasks without adding great mathematical complexity, therefore we will propose to create a mathematical model using Hermite spline interpolation to achieve the control given by (5) while keeping most of the parameters in (5).

6. Conclusion

We have reviewed some of the most significant work in coarticulation, and the development of our animation framework to accommodate the dynamic coarticulation. While our work is at its initial stages, we have made several advances that allow us to create performant lip-sync coarticulation, with the integration of our layered animation manager, which allows to blending MPEG-4 animation FAP animation with complex animations such as visemes and expressions, allowing these to be recorded. The XML specification permits the creation of visemes sets in our framework's editor and allows the shape of these to be modified at all times. Our framework generates dynamic lip-sync by using MaryTTS and Cohen-Massaro model (Massaro, 1993). Whilst this model is very generic and efficient, it needed however to create automatically several coarticulation parameters by aggregating information manually using test subjects. This is a laborious process that requires time, test subjects and equipment to record and analyse the results, which led us to devise the automation of the different parameters.

At the present time we have automated the calculation of the peak magnitude $\propto_{sp}$ used for the dominance function defined in (5) for each viseme set. In the future we will be looking to expand this automation process to include other parameters and we will modify the dominance functions to be used with more flexible interpolation curve functions to solve continuity issues and the introduction of silence visemes in the utterance. However, we have been able to show some promising results for some simple utterances and compared them with (King and Parent, 2005, Massaro, 1993, Pelachaud, 1991). We feel that the initial results allow us to be optimistic for the future creation of a resource efficient and realistic talking head which can support coarticulation and speech synchronization in MPEG-4 based facial animation.

References

Ardor3D [Online]. Available: http://ardor3d.com/ 2011].

FaceGen [Online]. Available: http://www.facegen.com/ 2014].

ALBRECHT, I., HABER, J. & SEIDEL, H.-P. 2002. Speech Synchronization for Physics-Based Facial Animation.

ALBRECHT, I., SCHRÖDER, M., HABER, J. & SEIDEL, H.-P. 2005. Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Special issue of Journal of Virtual Reality on Language, Speech & Gesture,* 8.

ANNOSOFT. *Annosoft viseme to phoneme set.* [Online]. Available: http://www.annosoft.com/phoneset.htm [Accessed 14-01-2013 2013].

BELL-BERTI, F. & HARRIS, K. 1979. Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America***,** 1268--1270.

BENGUEREL, A.-P. & PICHORA-FULLER, M. K. 1982. Coarticulation Effects in Lipreading. *J Speech Hear Res,* 25**,** 600-607.

BREGLER, C., COVELL, M. & SLANEY, M. 1997. Video Rewrite: driving visual speech with audio. *Proceedings of the 24th annual conference on Computer graphics and interactive techniques.* ACM Press/Addison-Wesley Publishing Co.

COHEN, M. M., MASSARO, D. W. & CLARK, R. 2002. Training a Talking Head. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces.* IEEE Computer Society.

COSI, P., FUSARO, A. & TISATO, G. 2003. LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. 127-132.

DENG, Z. & NOH, J. 2007. Computer Facial Animation: A Survey. *In:* DENG, Z. & NEUMANN, U. (eds.) *Data-Driven 3D Facial Animation.* Springer London.

EL RHALIBI, A., CARTER, C., COOPER, S., MERABTI, M. & PRICE, M. 2010. Charisma: High-performance Web-based MPEG-compliant animation framework. *Comput. Entertain.,* 8**,** 1-15.

GOFF, B. L. Automatic modeling of coarticulation in text-to-visual speech synthesis. *In:* KOKKINAKIS, G., FAKOTAKIS, N. & DERMATAS, E., eds. EUROSPEECH, 1997. ISCA.

KENT, R. D. & MINIFIE, F. D. 1977. Coarticulation in recent speech production models. *Journal of Phonetics,* 5**,** 115-135.

KING, S. A. & PARENT, R. E. 2005. Creating speech-synchronized animation. *IEEE transactions on visualization and computer graphics,* 11**,** 341-352.

KORAY, B. 2004. Xface: MPEG-4 based open source toolkit for 3D Facial Animation. *Proceedings of the working conference on Advanced visual interfaces.* Gallipoli, Italy: ACM.

LÖFQVIST, A. 1990. Speech as Audible Gestures. *In:* HARDCASTLE, W. J. & MARCHAL, A. (eds.) *Speech Production and Speech Modelling.* Kluwer Academic Publishers.

MASSARO, D. W. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press/ Bradford Books.

MASSARO, D. W. & COHEN, M. M. 1993. Modeling Coarticulation in Synthetic Visual Speech *Models and Techniques in Computer Animation.* Springer-Verlag.

MCGURK, H. & MACDONALD, J. 1976. Hearing lips and seeing voices. *Nature,* 264**,** 746-748.

OHMAN, S. E. G. 1967. Numerical Model of Coarticulation. *The Journal of the Acoustical Society of America,* 41**,** 310-320.

PANDZIC, I. S. & FORCHHEIMER, R. 2002. *MPEG-4 Facial Animation The Standard, Implementation and Applications*, John Wiley & Sons Ltd.

PASQUARIELLO, S. & PELACHAUD, C. 2001. Greta: A Simple Facial Animation Engine. *6th Online World Conference on Soft Computing in Industrial Applications, Session on Soft Computing for Intelligent 3D Agents*.

PELACHAUD, C. 1991. *Comunication and Coarticulation in facial Animation.* PhD, University of Pennylvania.

PELACHAUD, C. 2002. Visual Text-to-Speech. *MPEG4 Facial Animation - The standard, implementations and applications, Igor S. Pandzic, Robert Forchheimer (eds.), John Wiley & Sons*.

RUTH, C. & BARBARA, D. 1980. Hearing by eye. *Quarterly Journal of Experimental Psychology*.

SCHRÖDER, M. The German Text-to-Speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology, 2001. 365-377.

SOMASUNDARAM, A. 2006. *A Facial Animation Model for Expressive Audio-Visual Speech.* PhD, The Ohio State University.

SUMEDHA, K. & MAGNENAT-THALMANN, N. 2003. Visyllable Based Speech Animation. *Computer Graphics Forum,* 22.

TERRY, L. & KATSAGGELOS, A. K. A phone-viseme dynamic Bayesian network for audio-visual automatic speech recognition. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 8-11 Dec. 2008. 1-4.