Una Propuesta de Integración de Animación Facial y Voz Sintética

José F. Ferreira
Universidad de Los Andes, Departamento de Sistemas y Computación,
Bogotá D.C., Colombia
j-ferrei@uniandes.edu.co

y

Fernando De la Rosa
Universidad de Los Andes, Departamento de Sistemas y Computación
Bogotá D.C., Colombia
fde@uniandes.edu.co

Abstract

The work presented in this document consists of an exploration of modern computer facial animation techniques. The goal of our research is the analysis, evaluation and study of feasibility for integration of these techniques with existing text to speech translation tools. This integration provides the user a method of virtual characters creation that allows the generation of facial movements which are synchronized with the synthetic speech generated automatically by the text to speech engine. As a base for the research, the use of the frame provided by the compression and transmission of multimedia standard MPEG-4 was decided. This standard includes a specification of the concepts applicable to computer facial animation. The results obtained from the evaluated techniques about the generated animation quality are satisfactory and demonstrate the possibility of use of the generated virtual characters with voice in computer applications as an user interaction metaphor.
Keywords: Facial Animation, Computer Animation, MPEG-4, Synthetic Voice, Text To Speech Engines.

Resumen

El trabajo presentado en este documento consiste en una exploración de las técnicas de animación facial actuales. El objetivo de nuestra investigación es el análisis, evaluación y estudio de factibilidad de la integración de estas técnicas con herramientas existentes de *traducción de texto escrito a voz sintética*. Esta integración provee al usuario de un método de generación de personajes virtuales que permite la generación de movimientos faciales sincronizados con la voz sintética generada automáticamente por el motor de traducción de texto a voz sintética. Para el desarrollo de la investigación se eligió la utilización del marco provisto por el estándar de compresión y transmisión de multimedia MPEG-4 que incluye una especificación de los conceptos aplicables a la animación facial. Los resultados obtenidos de la evaluación de las técnicas estudiadas sobre la calidad de la animación generada son satisfactorios y demuestran la posibilidad del uso de personajes virtuales animados, con voz, en aplicaciones computacionales como metáfora de interacción con el usuario.
Palabras claves: Animación Facial, Animación por computador, MPEG-4, Voz Sintética, Motores de traducción de Texto a Voz.

1. Introducción

Uno de los factores más importantes en el crecimiento de la computación en las últimas dos décadas ha sido la introducción y desarrollo de técnicas de digitalización de medios audiovisuales. Con el aprovechamiento de las capacidades multimedia de los computadores actuales se ha logrado transmitir, con efectividad y de manera dinámica, diferentes tipos de información de manera simultánea captando una mayor atención del usuario hacia el mensaje.
Sin duda alguna, la posibilidad de usar sonidos, imágenes, videos, animaciones y otras formas de comunicación en su forma digital, le ha dado un gran impulso a la industria informática de nuestro tiempo. Particularmente, en el área de la informática gráfica muchos de los esfuerzos realizados han sido destinados a lograr una representación realista del mundo real a través del uso de información digital.
Uno de los elementos fundamentales de la representación del mundo real en un ambiente digital es la representación de nosotros mismos; por esta razón, una de las áreas de investigación más activa en el campo de la computación gráfica es la animación facial de modelos geométricos para la representación de rostros.
Las técnicas desarrolladas por estas investigaciones han logrado ser aplicadas con éxito en diversas áreas tales como entretenimiento y educación. Hoy en día, este tipo de avances, junto al mejoramiento sustancial del hardware gráfico disponible en el mercado, ha permitido el reciente incremento de la utilización de personajes virtuales realistas en aplicaciones de diferente tipo como películas cinematográficas, software educativo, publicidad, entretenimiento, ambientes virtuales (chat, centros comerciales,…), etc.
La investigación presentada en este trabajo explora los conceptos en los que se basan las técnicas de animación facial actuales (representación geométrica, interpolación, reproducción de fonemas,…). A partir del conocimiento de éstos, se evalúa la factibilidad de la integración de herramientas existentes de traducción de texto escrito a voz sintética con el fin de proveer al usuario de un método de generación de personajes virtuales que, además de tener una voz propia, tengan un movimiento facial sincronizado con los sonidos generados automáticamente.
El desarrollo de cualquier aplicación resulta mucho más útil cuando es compatible con alguno de los estándares definidos en el mercado. Por este motivo, nuestra decisión fue el realizar un estudio de las características definidas en el estándar MPEG-4 para el soporte de la animación facial.
El presente artículo está organizado de la siguiente manera: la sección 2 presenta los aspectos del estándar MPEG-4 para animación facial. En la sección 3 se exponen la problemática de deformación geométrica facial y su sincronización con la voz sintética. A continuación, se describe la arquitectura de la solución propuesta y una aplicación computacional de la misma. Finalmente se encuentran los trabajos futuros y las conclusiones.

2. El Estándar MPEG-4 para Animación Facial

MPEG-4 define en detalle los parámetros para la definición y animación de modelos faciales [11, 12]. Por un lado, los parámetros para la definición facial (*FDP: Facial Definition Parameters*) permiten una especificación completa de la forma, tamaño y textura del modelo de la cara. Por otro lado, los parámetros de animación facial (*FAP: Facial Animation Parameters*) hacen posible la representación de expresiones faciales y visemas (i.e. conjunto de apariencias visuales relacionadas con un fonema; también llamados "*fonemas visuales*") mediante la manipulación de puntos característicos del modelo geométrico. Los *FAPs* son expresados en términos de *FAPUs* (*Facial Animation Parameter Units*) que consisten en la representación de la distancia entre los diferentes puntos característicos definidos para conservar una proporcionalidad adecuada al tamaño del modelo geométrico.
En el caso de la definición de los modelos faciales, los *FDPs* son usados para personalizar un modelo genérico (Ver Figura 1). Los campos definidos para una definición completa del modelo son las coordenadas de los puntos característicos, las coordenadas de la textura aplicada al modelo y el comportamiento de los parámetros de animación dentro de la escena [9, 13, 23].
El estándar MPEG-4 utiliza métodos de animación basados en parametrizaciones de la posición de un conjunto de puntos característicos predefinidos. En la especificación del estándar MPEG-4 se definen los ya mencionados *FAPs* que consisten, en la mayoría de los casos, de un desplazamiento unidireccional o bidireccional del punto característico asociado.
La definición de estos parámetros se basa en el estudio de las acciones musculares del rostro humano y comprende 68 parámetros divididos en dos grupos: parámetros de animación de bajo y de alto nivel [2, 11, 13]. Los *FAPs* de bajo nivel están relacionados con los puntos característicos definidos por los *FDPs* y generalmente son una medida de desplazamiento relativo de estos puntos dentro del modelo. Los *FAPs* de alto nivel comprenden las expresiones y los visemas. Estos dos últimos *FAPs* pueden mover al mismo tiempo un conjunto de puntos característicos modificando por completo el gesto del rostro. En particular, los visemas manipulan la postura y la posición de los labios para la

representación visual de la vocalización de un fonema [2, 5]. En el caso de las expresiones, se manipulan varios subconjuntos de *FAPs* de bajo nivel para obtener la representación de una expresión facial.
Debido a que estos parámetros deben poder ser usados en modelos geométricos de diferentes tamaños y características se definen también las *FAPUs*. Estas unidades son definidas como fracciones de distancias con respecto a características claves del rostro en un estado neutral. Los valores de los *FAPs* están dados en términos de *FAPUs* para lograr un desplazamiento adecuado en modelos de diferentes tamaños [12,13].

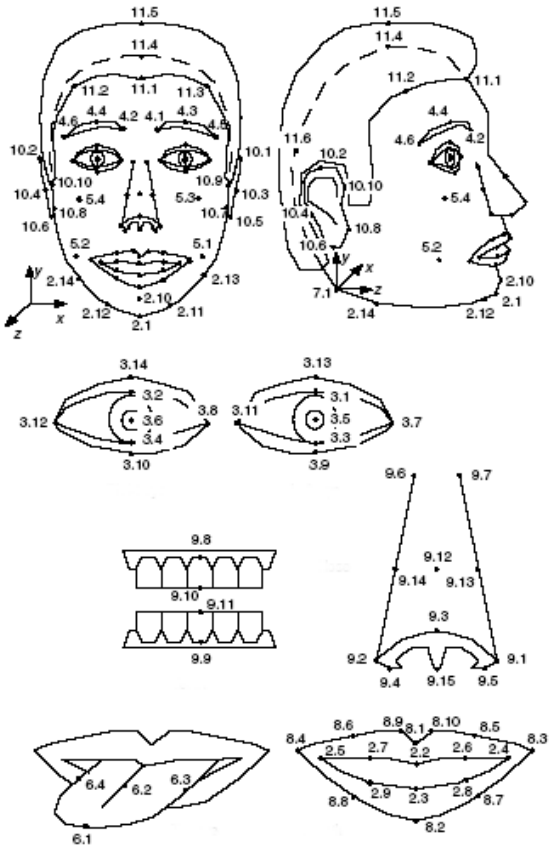

Figura No. 1. MPEG-4 ha definido un conjunto de 84 puntos característicos usados para calibrar y animar un rostro. Los puntos son clasificados de acuerdo a la región del rostro a la que pertenecen. [17]

Para producir un rostro animado compatible, MPEG-4 no especifica ninguna condición sobre el modelo utilizado a excepción que los puntos característicos definidos para el rostro sean desplazados de acuerdo a las magnitudes definidas por los *FAPUs* particulares del modelo geométrico durante la interpretación de los *FAPs* de bajo nivel.

3. Principales problemáticas identificadas

El principal problema que resuelve nuestra investigación es el análisis, diseño, implementación y pruebas de un sistema computacional capaz de crear un personaje virtual realista (representado por un modelo geométrico facial 3D) con la capacidad de realizar movimientos faciales sincronizados con una voz sintética generada a partir de un texto escrito.
De manera adicional, el sistema debe permitir que el personaje virtual creado sea una representación realista de un rostro de una persona real de manera que sea más natural la utilización de la aplicación para el usuario final.
A continuación se explica cada uno de los problemas que requieren ser tratados para lograr una solución completa, indicando las soluciones específicas propuestas para cada uno de ellos en la aplicación de prueba.

3.1 Deformación del modelo facial usando parámetros MPEG-4

La deformación del modelo facial es quizás el problema de mayor importancia en el alcance de esta investigación. El modelo facial debe ser deformado para generar las expresiones faciales declaradas por los parámetros de animación. En la solución de este subproblema se debieron identificar aquellas técnicas de animación facial que permiten el desarrollo

de una aplicación compatible con MPEG-4. Con el propósito de determinar qué enfoque existente de los usados en otras investigaciones resulta más adecuado para lograr el grado de compatibilidad deseado fue necesario relacionar los conceptos investigados sobre la animación facial con los principios establecidos por MPEG-4 para el manejo de este tipo de animaciones. Para lograr una deformación facial adecuada y lo más realista posible, es necesario estudiar y resolver los siguientes subproblemas específicos:

- La representación del modelo geométrico: Para lograr que un método de representación geométrica de la cabeza humana sea compatible con MPEG-4 basta con que el modelo tenga un vértice asociado con cada punto característico definido en MPEG-4. No es necesario incluir todos los puntos estandarizados dentro de una aplicación para lograr compatibilidad con MPEG-4 [12, 23]. Adicionalmente, para cada punto característico se deben definir sus respectivas coordenadas de textura.

- La selección interactiva de los puntos a deformar: Con el objetivo de solucionar la problemática de selección de puntos característicos, MPEG-4 introduce el concepto de *FDPs*. Estos parámetros pueden ser usados ya sea para modificar la forma y apariencia de un modelo o para codificar la información necesaria en la transmisión de un modelo completo junto con los criterios que deben ser usados para animarlo. Los *FDPs* son usados con el fin de cambiar la ubicación relativa de los puntos característicos definidos por el estándar MPEG-4 [12]. Sin embargo, dicho estándar no define un método específico de generación y codificación de *FAPs* ni *FDPs* [1], por lo que cada aplicación puede definir un método interno de representación de estos parámetros siempre y cuando la transmisión de esta información hacia el exterior de la aplicación siga las normas semánticas y sintácticas acordes a MPEG-4 [12]. Para el caso específico de nuestra aplicación de prueba, el soporte de este tipo de transmisión externa no fue necesario, por lo que el único aspecto a resolver consiste en la definición de la representación interna de la información necesaria para la animación del modelo facial en la aplicación.

- La deformación de grupos de vértices: Cuando el modelo es modificado durante una animación es necesario que los vértices cercanos a aquel que representa un punto característico, sean también movidos consistentemente con el fin de lograr un mayor realismo [13].

- La interpolación de los parámetros de animación usados: Los *FAPs* que controlan la animación son valores fijos definidos para un instante de tiempo preciso, y por lo tanto una animación que considere únicamente los movimientos para estos instantes de tiempo se vería poco realista (i.e. se apreciarían cambios bruscos entre los instantes definidos). Por este motivo es necesaria la inclusión de técnicas de interpolación de animación para los instantes de tiempo que se encuentran entre dos cuadros de animación definidos por dos *FAPs* consecutivos.

En resumen, para ofrecer una solución completa al problema de animación facial, es necesaria la inclusión de técnicas adecuadas de selección de puntos en un modelo facial 3D que permitan al usuario escoger los *FDPs* de una representación geométrica 3D compatible con MPEG-4, así como una forma de determinar el conjunto de vértices que se ven afectados debido a la deformación de un punto característico y que es definida por los *FAPs*. También es necesario definir los métodos de interpolación de animación de cuadros para suavizar los movimientos determinados por los *FAPs* que son parámetros discretos de animación, razón por la cual no definen el estado del modelo en todos los instantes posibles de tiempo.

3.2 Sincronización de voz sintética y animación

Una característica fundamental de nuestra investigación es la sincronización de la animación con la voz sintética generada por los motores de traducción a partir de un texto escrito. Dicha voz sintética debe corresponder con los movimientos realizados por los parámetros de animación facial para preservar un nivel de realismo aceptable. Debido a que los módulos de animación y síntesis de voz son independientes, el control del tiempo de ejecución debe ser resuelto en la etapa previa a la generación de las ondas de audio a partir del texto de entrada.

Los sistemas de síntesis de voz generada a partir de un texto deben proveer al sistema de animación facial tanto de la información sobre los fonemas que deben ser representados como su aparición en el tiempo en la onda de audio previamente generada. En el caso de la especificación definida por el estándar MPEG-4, aunque se contempla la integración de motores TTS (*Text To Speech*) con el resto de las tecnologías a través de una interfaz genérica llamada TTSI (*Text To Speech Interface*) [4], no se aclara la utilización de información de sincronización entre el audio y la información de animación. Por este motivo, es necesario definir una arquitectura externa que genere un flujo sincronizado de las dos fuentes, el motor TTS y la interfaz de animación del modelo. Ostermann *et al.* [14] proponen una arquitectura de sincronización basada en *FAPs* usando un mecanismo de marcas de tiempo (llamadas *"bookmarks"*) y una función de interpolación de puntos para determinar el estado del modelo entre las marcas de tiempo. Para la sincronización de las salidas de audio y animación, se utiliza la información de los tiempos de reproducción de los fonemas generados por el motor de traducción para calcular los cuadros de la animación del modelo facial.

La propuesta de Ostermann *et al.* [14] se usó como base para la definición de la arquitectura de la aplicación de prueba desarrollada durante nuestra investigación en cuanto a las ideas expuestas sobre el proceso de transformación de la información proveniente del motor de traducción TTS en parámetros MPEG-4 utilizables por el módulo encargado de la generación de la animación.

4. Implementación de la solución en la Aplicación de Prueba

Con el fin de validar la aplicabilidad de las técnicas estudiadas en un caso real y de evaluar las soluciones encontradas durante la investigación a las problemáticas definidas en la sección anterior se implementó una aplicación de prueba en lenguaje C/C++ usando el API de programación OpenGL [22]. A continuación se exponen algunos detalles de la estructura del sistema desarrollado.

4.1 Arquitectura general del sistema

En la Figura No. 2 se muestra el esquema general de la arquitectura del sistema propuesto. En el esquema se pueden identificar claramente tres módulos:

- Módulo de Definición Facial encargado de la definición del modelo geométrico a través del soporte del formato OBJ y la representación de los puntos característicos compatibles con MPEG-4 que serán definidos de manera interactiva por el usuario (Figura No. 3).
- Módulo de Animación Facial, que se ocupa del manejo de los parámetros de animación y su correspondiente efecto en el estado de los vértices del modelo incluyendo los aspectos de interpolación entre los cuadros de animación definidos por los parámetros y de deformación de vértices vecinos a un punto característico movido (Figura No. 4).
- Módulo de Decodificación de Texto y Síntesis de Voz, que tiene como función principal controlar la generación de los parámetros de animación de acuerdo a la información proveniente de la decodificación del texto escrito (Figura No. 5). El proceso de decodificación y síntesis de voz se realizará con el Festival Speech Synthesis System [20].

Al interior de la aplicación (Figura No. 2), el componente de sincronización recibe por separado los parámetros de animación y los de voz sintética y reproduce las señales video y audio de manera simultánea.

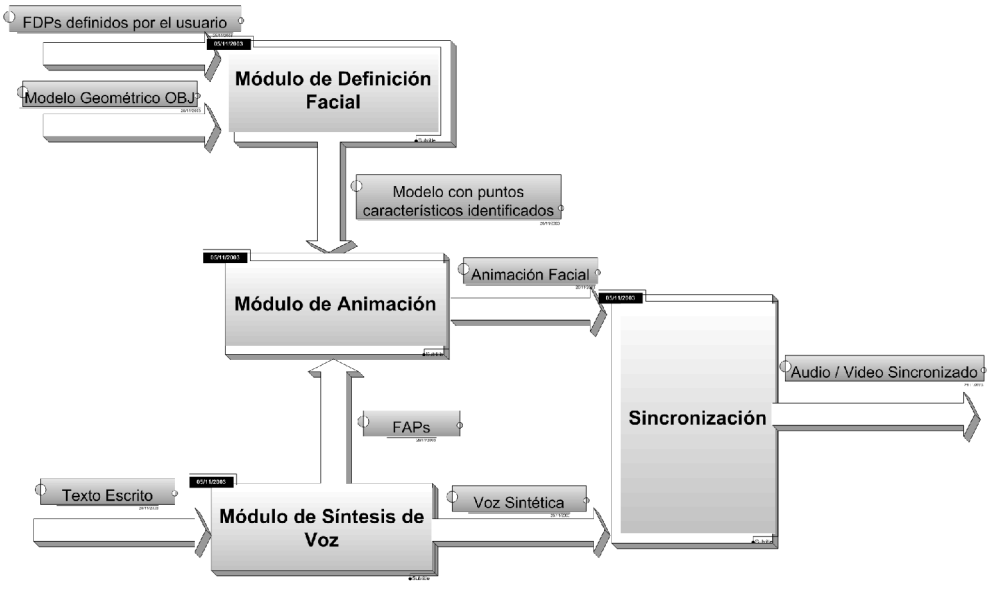

Figura No. 2. Esquema general de la arquitectura del sistema propuesto

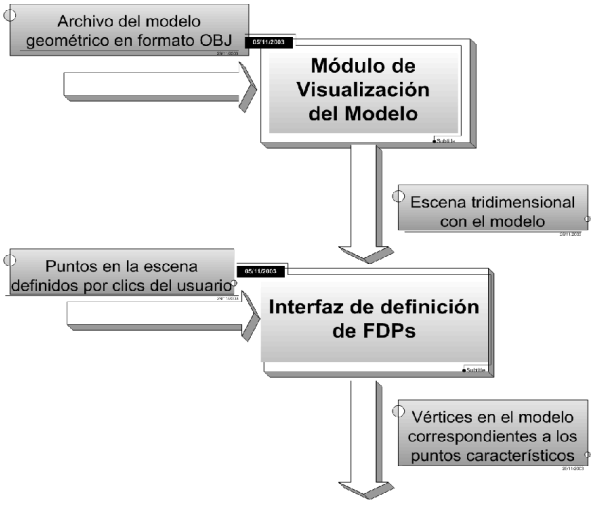

Figura No. 3. Módulo de definición del modelo facial.

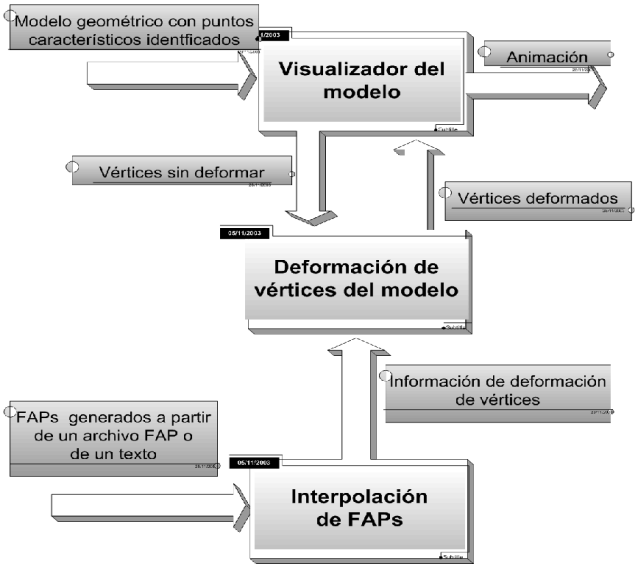

Figura No. 4. Módulo de animación facial.

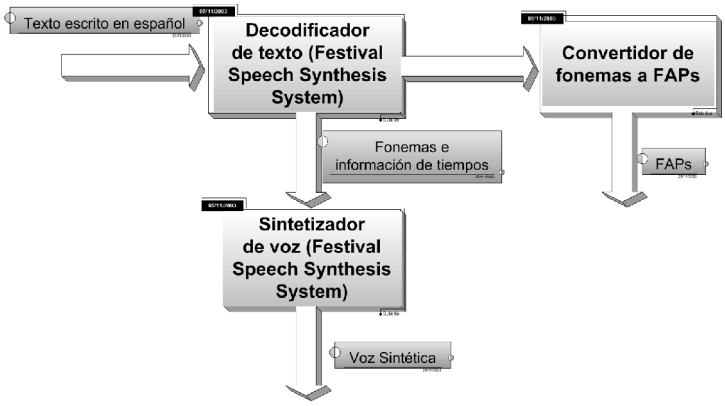

Figura No. 5. Módulo de decodificación de texto y síntesis de voz.

4.2 Despliegue del modelo y selección de puntos característicos

Para la inclusión de la selección de puntos característicos en la aplicación de prueba se utilizó el mecanismo de *color coding* [16]. Esta alternativa se eligió debido a la facilidad de implementación y a la independencia que ofrece de la posición del observador y del estado del modelo que en los otros métodos evaluados añaden cierto grado de complejidad y dificultan una detección adecuada del vértice seleccionado.

El mecanismo de *color coding* se basa en la utilización de un código de colores diferentes para cada elemento seleccionable del modelo. En el caso específico de la aplicación de prueba se utiliza una codificación especial sobre los valores RGB del color de los vértices de manera que cuando el usuario hace *clic* con el ratón sobre uno de los vértices del modelo, se obtienen los valores RGB del píxel y se transforman en el identificador del vértice en el modelo de representación interna.

4.3 Deformación del modelo geométrico

El control de la animación generada es establecido por los parámetros *FAP* como lo establece el estándar MPEG-4. En la aplicación de prueba se definen las estructuras que guardan la información proveniente de archivos *FAP* o del análisis de la información de síntesis de voz. Posteriormente, esta información es usada para determinar los puntos característicos a deformar en el modelo y lograr la animación del modelo facial.

Se decidió aplicar el mecanismo más sencillo posible de deformación de vértices vecinos y evaluar el desempeño y nivel de realismo de la animación lograda. Por esta razón, la aplicación utiliza un método de deformación constante de grupos predefinidos de vértices para lograr la animación del modelo.

El proceso de deformación implementado consta, básicamente, de dos etapas:

- Inicialización: Se definen los grupos de vértices a partir de los puntos característicos definidos por el usuario sobre el modelo. No todos los puntos característicos determinan la creación de un grupo de vértices, sólo aquellos que se consideran esenciales para una correcta apariencia de los movimientos básicos del rostro. El detalle de los grupos de labios a partir de subgrupos se determinó por el efecto que éstos tienen en la apariencia realista del movimiento involucrado en la gesticulación de visemas. Adicionalmente se calculan las unidades *FAPU* correspondientes al modelo a partir de los puntos característicos definidos por el estándar MPEG-4.

- Deformación en tiempo real: El conjunto de vértices asociado a un punto característico seleccionado se deforma en una proporción igual al movimiento (i.e. cambio) de este punto característico. Este método de deformación, contrario a lo que se podría pensar, logra un realismo de movimiento aceptable (en deformaciones con magnitudes relativamente pequeñas) unido a un desempeño superior comparado con el de otras alternativas evaluadas (métodos de *warping* [15,18], *free form deformations* [6, 7, 10, 21], deformaciones basadas en los puntos característicos [8]) que involucran cálculos más intensivos en términos computacionales.

4.4 Interpolación de parámetros de animación

La definición de cada cuadro de animación mediante los parámetros de animación no es suficiente para la obtención de una animación realista debido a que entre dos cuadros consecutivos de animación no se tiene información sobre la posición de los vértices del modelo. La solución a este problema consiste en la inclusión de técnicas de interpolación temporal o *in-betweening* de los cuadros de animación. Cuando la aplicación intenta mostrar el estado del modelo facial en un tiempo situado entre dos cuadros de animación se calcula un valor intermedio, entre los parámetros del cuadro inicial y los parámetros del cuadro consecutivo, que depende del tiempo transcurrido en la animación.

Después de evaluar las diferentes posibilidades para la implementación de interpolación temporal entre dos cuadros de animación se decidió utilizar una función lineal definida para el cambio de posición de los vértices en el tiempo. Este cálculo se realiza para cada ciclo de visualización *(i.e. rendering)* de la escena, usando como parámetro de la función de interpolación el tiempo transcurrido desde el inicio de la reproducción de la animación.
La elección de esta alternativa se basó fundamentalmente en la intención de liberar a la aplicación de prueba de cálculos más complejos obteniendo de todas formas una buena solución. El análisis de funciones de interpolación más complejas determinó que éstas no introducían una mejora significativa en la calidad de la animación generada.

4.5 Sincronización de las salidas del sistema

Desde el inicio de la investigación se determinó que para la inclusión de un módulo de decodificación y síntesis de voz se debería utilizar software ya existente en el mercado debido a la alta complejidad de las posibles soluciones a esta problemática.
Por ésta razón uno de los estudios realizados durante nuestra investigación consistió en la evaluación de los diferentes sistemas de decodificación y síntesis disponibles para determinar cuál podría ser utilizado.
Como resultado se encontró que trabajos relacionados con el que se presenta en este documento [3,19] han utilizado con éxito las funcionalidades ofrecidas por el Festival Speech Synthesis System [20] para la traducción de texto en fonemas con la ventaja sobre otras soluciones de que esta aplicación está disponible de manera gratuita para fines investigativos.
El Festival Speech Synthesis System soporta el español y provee un API en C/C++ que tiene la capacidad de recibir como parámetros de entrada un texto y generar una trascripción equivalente de los fonemas contenidos en él junto con información de la duración de cada uno de ellos. Esta información es analizada por un lado para producir la onda de audio que será reproducida por la aplicación e igualmente para la traducción de fonemas y marcas de tiempo a *FAPs* lo que asegura la sincronización del audio y la animación.
La traducción de fonemas a *FAPs* se basa en el siguiente método. Debido a que la animación tiene una velocidad predefinida dada en cuadros por segundo y la información de inicio de los fonemas y su duración está disponible, se realiza una traducción de los fonemas a parámetros *FAP* de alto nivel, codificando el visema correspondiente en el cuadro de la animación más cercano al tiempo especificado por la información de síntesis. Debido a que esta aproximación introduce un pequeño desfase en la sincronización de la animación y la voz, se realiza un proceso de interpolación adicional entre dos cuadros que contienen visemas para generar la información intermedia en términos de *FAPs* de bajo nivel.
Otro de los puntos a resolver para la solución de esta problemática al interior de la aplicación de prueba, fue el de lograr una correspondencia adecuada entre los visemas definidos por el estándar MPEG-4 y los fonemas resultado del análisis del texto de entrada en Español realizado por el motor de traducción de texto de Festival. La propuesta de traducción especificada se puede ver en la Tabla No. 1.

VISEMAS MPEG-4		***FONEMAS FESTIVAL ESPAÑOL***
0	Ninguno	#
1	p, b, m	P, b, m
2	f, v	F
3	T, D	Ninguno
4	t, d	t, d
5	k, g	k, g
6	tS, dZ, S	x, ch, ll
7	s, z	s, th (z)
8	n, l	n, ny (ñ), l
9	R	r, rr
10	A	a, a1
11	E	e, e1
12	I	i, i0, i1
13	Q	o, o1
14	U	u, u0, u1

Tabla No. 1. Correspondencia entre los visemas definidos por MPEG-4 y los fonemas generados por Festival a partir del texto de entrada en lenguaje Español.

Con la integración de las soluciones propuestas a las problemáticas definidas se consigue obtener una aplicación con las funcionalidades esperadas. El esquema general de funcionamiento de la aplicación y el proceso de transformación de la información es mostrado en la Figura No. 6. Adicionalmente, la Figura No. 7 muestra la aplicación implementada basándose en las técnicas expuestas en este artículo.

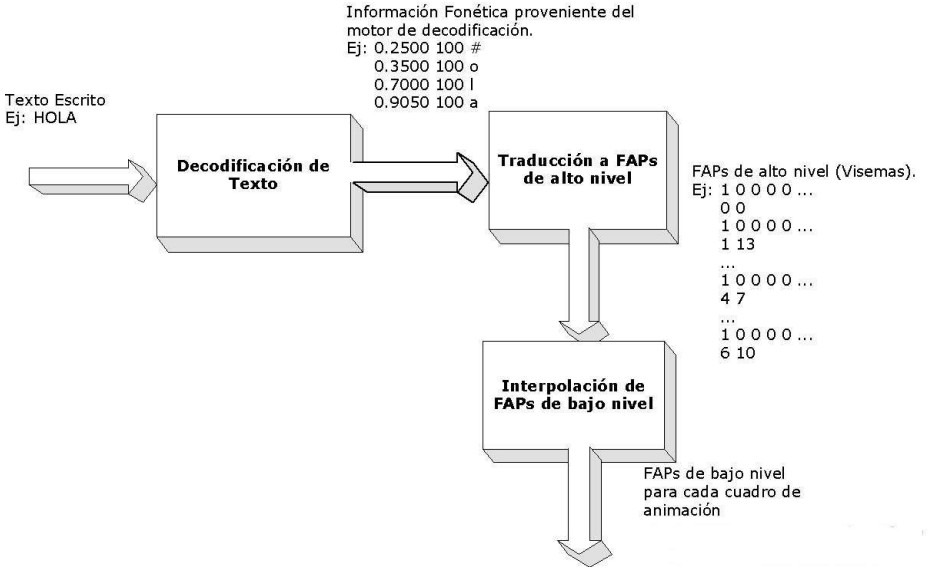

Figura No. 6. Proceso de sincronización de las salidas de audio y animación

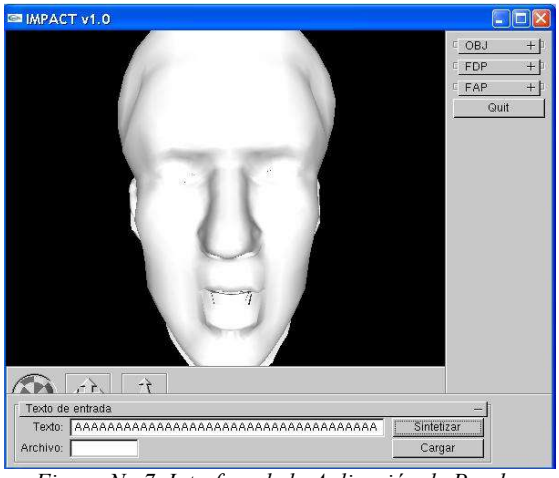

Figura No.7. Interfase de la Aplicación de Prueba

5. TRABAJO FUTURO

Nuestra investigación cumplió con el objetivo de proponer una solución al problema de integrar la Animación Facial con Voz Sintética. Sin embargo, existe una gran variedad de estudios que podrían desarrollarse para dar un tratamiento más amplio al tema:

5.1 Posibles aplicaciones de los conceptos desarrollados

Uno de los temas de investigación que probablemente tendría mayor importancia sería la definición de aplicaciones que utilicen las funcionalidades definidas en la aplicación de prueba para fines específicos. Entre los tipos de aplicaciones que se podrían desarrollar se encuentran:

- Ambientes Virtuales Compartidos (accesibles a través de Internet) como por ejemplo una aplicación de pseudo-teleconferencia o "chat visual" que a través de la utilización de personajes virtuales facilite el diálogo entre dos o

más personas al usar información tanto visual como sonora transmitida en términos de parámetros de animación y procesada de manera local para la generación de la animación.

- Interfaces antropomórficas que agreguen un componente adicional de facilidad de uso y de interacción con el usuario mediante la representación de un personaje virtual que sea el encargado de comunicar los diferentes eventos al usuario. Este tipo de interfaces sería muy útil en aplicaciones educativas o informativas en los que los usuarios carezcan de un conocimiento extenso de la manipulación del sistema o sean personas con discapacidades físicas que no les permitan una interacción tradicional con el sistema.
- Aplicaciones de generación de animaciones para uso en juegos u otros medios audiovisuales que faciliten la introducción de personajes virtuales resolviendo automáticamente la problemática de sincronización de labios con las voces *(lip sync)*.
- Aplicaciones variadas de entretenimiento en las que una de las formas de interacción con el sistema sea la introducción de texto. Por ejemplo, una aplicación que soporte la visualización de una historia en la que el usuario sea partícipe y modifique o determine los sucesos que ocurren durante el desarrollo de la trama.

5.2 Extensiones a las funcionalidades de la aplicación de prueba

El trabajo realizado durante la investigación deja abiertas algunas alternativas para un desarrollo más complejo de algunas de las características de la aplicación de prueba:

- Implementación de un módulo propio de obtención de los modelos faciales a partir de fotografías que permitiría un mejor control sobre las propiedades del modelo mediante la introducción de un modelo genérico con puntos característicos predefinidos que sirva como base para la creación de los modelos personalizados. También sería interesante estudiar la factibilidad de desarrollo de módulos de obtención de modelos a partir de otras fuentes como el vídeo y la detección automática de los puntos característicos del rostro.
- Implementación de animación corporal (también especificada en el estándar MPEG-4) de manera análoga a la animación facial presentada.
- Integración con otros sistemas que amplíen la gama de usos de la aplicación. Dentro de este tipo de sistemas se encuentran los sistemas de reconocimiento de voz que integrados con los resultados actuales permitiría generar directamente una animación facial a partir de la voz de una persona. Igualmente la integración con sistemas de inteligencia artificial basados en agentes (i.e. entidades autónomas) capaces de mantener una conversación con un usuario mediante la generación de textos de entrada.
- Soporte de formatos de definición de modelos geométricos variados (actualmente hay soporte de modelos OBJ).
- Generación de salidas en formatos de animación o vídeo usados de manera amplia, especialmente los definidos por el estándar MPEG-4. Dichas salidas podrían ser visualizadas en otras aplicaciones que soporten dichos formatos.
- Integración de otras técnicas de deformación para mejorar el nivel actual de realismo de las animaciones generadas.

6. Conclusiones

La primera etapa de nuestra investigación consistió en la comprensión de los conceptos en los que se fundamenta los temas de animación facial y síntesis de voz. Posteriormente, investigamos y analizamos los métodos necesarios para desarrollar un sistema computacional que resuelven e integran las dos problemáticas. La implementación de los métodos finalmente escogidos se integraron en la aplicación de prueba, resultado de este trabajo.

Las evaluaciones practicadas hasta el momento cumplen los objetivos inicialmente previstos. Sin embargo, el nivel de los resultados puede mejorar mediante su aplicación en áreas específicas (ambientes virtuales compartidos, entretenimiento, educación, publicidad,...).

Uno de los factores más importantes dentro del desarrollo de la investigación fue la elección del estándar MPEG-4 ya que éste provee la definición precisa de un marco que resulta apropiado para la construcción de aplicaciones de animación computacional y, de cierta manera, establece un esquema general que resulta muy útil para la definición de una arquitectura apropiada. Adicionalmente, MPEG-4 suministra una forma de controlar animaciones relativamente complejas (como lo son las animaciones faciales) a partir de un flujo simple de datos. Dicho flujo permite una definición simple de interacción entre los diferentes módulos de la aplicación y una integración sencilla con el software de síntesis de voz que, en principio, es uno de los objetivos primordiales de este proyecto.

Otro de los factores relevantes en la problemática estudiada es la elección del motor de decodificación y síntesis de voz. En nuestro trabajo, Festival Speech Synthesis System resultó ser una buena elección debido a la variedad de opciones de control de las salidas de información fonética y voz que genera. Esto hace posible el proceso de sincronización animación – voz sintética. Vale la pena mencionar que tanto Festival como otros motores del mismo tipo no generan (por ahora) voz humana natural.

Referencias

[1] Antunes-Abrantes, G. y Pereira, F. MPEG-4 Facial Animation Technology: Survey, Implementation and Results, *IEEE Transactions on Circuits and Systems*. Vol. 9, No. 2, (Marzo 1999), pp. 290-305.

[2] Eptamedia, Facial Animation Background, *http://www.eptamedia.com/en-doc/en-faceanim-c-background.htm*. (Revisado en Noviembre 2003)

[3] Huynh, H.Q. A Facial Animation Markup Language (FAML) for the Scripting of a Talking Head. Curtin University of Technology, (2000).

[4] ISO/IEC JTC1/WG11 N2201, Texto de la especificación ISO/IEC FCD 14496-1: Systems. 1998.

[5] Joslin, C., Molet, T., Magnenat-Thalmann, N. Distributed Virtual Reality Systems. MIRALab, CUI, University of Geneva, 2001.

[6] Kalra , P., Mangili, A., Magnenat-Thalmann, N., Thalmann D. 3D Interactive Free Form Deformations for Facial Expressions. MIRALab, University of Geneva, 1991.

[7] Kalra, P., Mangili, A., Magnenat-Thalmann, N., Thalmann, D. Simulation of Facial Muscle Actions Based on Rational Free Form Deformations. *Computer Graphics Forum*, Vol. 11, No. 3, (1992), pp. 59-69

[8] Kshirsagar, S., Garchery, S., Magnenat-Thalmann, N. Feature Point Based Mesh Deformation Applied to MPEG-4 Facial Animation. *Proceedings Deform'2000*; (2000).

[9] Lavaggetto, F. & Pockaj, R. The Facial Animation Engine: Towards a High-Level Interface for the Design of MPEG-4 Compliant Animated Faces. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 9, No. 2, (March 1999), pp.277-289

[10] Martin, Suzanne. Free – form Deformation. *http://www.cs.wpi.edu/~matt/courses/cs563/talks/smartin/ffdeform.html* (Revisado en Agosto 2003)

[11] Moving Picture Experts Group, ISO/IEC 14496 – MPEG-4 International Standard, *www.cselt.it/mpeg/*. (Revisado en Agosto 2003)

[12] MPEG Video & SNHC, Texto de la especificación ISO/IEC FDIS 14496-2: Visual.

[13] Ostermann, J. Face Animation in MPEG-4. *MPEG-4 Facial Animation (I.S. Pandzic and R. Forchheimer, Eds.)*, p.p.17-56, Chichester, U.K., John Wiley \& Sons, (2002).

[14] Ostermann, J., Beutnagel, M., Fischer, A., and Wang, Y. Integration of Talking Heads and Text to Speech Synthesizers for Visual TTS. *International Conference on Speech and Language Processing*. Sydney – Australia, (1998), p. 297-300

[15] Parent, R. A System for Generating Three-Dimensional Data for Computer Graphics ,Ph.D. Dissertation, Ohio State University, (1977).

[16] Picking Tutorial OpenGL. http://www.lighthouse3d.com/opengl/picking/.(Revisado en Agosto 2003)

[17] Pockaj, R., Baudino, M.,Corte, F., Ambrosini, L., Costa M., Bonamico C. The Facial Animation Engine Demo. *http://www-dsp.com.dist.unige.it/~pok/RESEARCH/MPEG/fae.htm*. (Revisado en Octubre 2003)

[18] Porcher, L. Animacao por Computador – Deformacao. Instituto de Informática UFGRS. *http://www.inf.ufrgs.br/~nedel* (Revisado en Septiembre 2003)

[19] Stallo, J. Simulating emotional speech for a talking head. Curtin University of Technology, (2000).

[20] The Centre for Speech Technology Research - University of Edinburgh. Festival Speech Synthesis System. *http://www.cstr.ed.ac.uk/projects/festival/.* (Revisado en Marzo 2004)

[21] Uribe, D. Uso de Humanoides para Comercio Electrónico. Universidad de Los Andes, (2002).

[22] Woo, M. OpenGL programming guide : the official guide to learning OpenGL, version 1.2. 3rd ed. Boston, Mass. : Addison Wesley , c1999.

[23] Won-Sook, L., Escher, M., Sannier G., Magnenat-Thalmann, N. MPEG-4 Compatible Faces from Orthogonal Photos. Proceedings CA99 (International Conference on Computer Animation), Geneva, Switzerland. May 26-29, 1999, pp.186-194.