Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Animación de Expresiones Faciales y del Movimiento de Labios en el Habla del Español en un Ambiente Tridimensional

T e s i s

que como requisito parcial para obtener el título de:

Ingeniero en Ciencias de la Computación

Presenta:

Héctor Adrián Díaz Furlong

Asesor de Tesis:

Dr. Manuel Martín Ortiz

F.C.C
B.U.A.P

Puebla, Pue. Agosto 2009

Animación de Expresiones Faciales y del Movimiento de Labios en el Habla del Español en un Ambiente Tridimensional

Tesis

Presentada a la Facultad de Ciencias de la Computación
como requisito parcial para la obtención del grado de

Ingeniero en Ciencias de la Computación

por

Héctor Adrián Díaz Furlong

en la

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

Agosto 2009

Firma del autor .....
Facultad de Ciencias de la Computación
28 de agosto del 2009

Aprobada por .....
Manuel Martín Ortiz
Doctor en Física
Asesor de Tesis

Aprobada por .....
Abraham Sánchez López
Doctor en Ciencias de la Computación y Robótica
Jurado

Aprobada por .....
Iván Olmos Pineda
Doctor en Ciencias de la Computación
Jurado

Aprobada por .....
Consejo de Ciencia y Tecnología del Estado de Puebla
CONCYTEP

Índice general

Animación de Expresiones Faciales y del Movimiento de Labios en el Habla del Español en un Ambiente Tridimensional

por

Héctor Adrián Díaz Furlong

Presentada a la Facultad de Ciencias de la Computación
el 28 de agosto del 2009, como requisito parcial para obtener el grado de
Ingeniero en Ciencias de la Computación

Resumen

En este trabajo se desarrolló un sistema de animación facial en 3D utilizando la técnica de metamorfosis con interpolación lineal. Se cuenta con un conjunto inicial de modelos 3D que contiene distintas expresiones faciales y visemas (fonemas visuales). La animación se realiza interpolando en el tiempo las posiciones de los vértices de un modelo fuente y un modelo destino. El sistema recibe como entrada un texto en español y genera la animación del rostro de tal manera que se logra la apariencia de que está pronunciando las palabras del texto. Para ello, el texto pasa por un proceso de silabeo, ya que fonológicamente los segmentos (consonantes y vocales) se agrupan en sílabas. Una vez que se tienen las sílabas, éstas son descompuestas en fonemas, que son proyectados a su conjunto de visemas correspondiente para producir la animación. El sistema puede emplearse como auxiliar en la terapia del lenguaje para niños con hipoacusia, ayudando a fortalecer sus habilidades de lectura labiofacial y permitiéndoles adquirir conocimientos del lenguaje a través de la vista.

Asesor de Tesis: Manuel Martín Ortiz
Título: Doctor en Física

Jurado: Abraham Sánchez López
Título: Doctor en Ciencias de la Computación y Robótica

Jurado: Iván Olmos Pineda
Título: Doctor en Ciencias de la Computación

Agradecimientos

Agradezco al CONCYTEP por el apoyo de beca que me brindó durante la realización de este trabajo.

A mis padres, Elizabeth Furlong Acevedo y Alfonso Díaz Cárdenas, quienes me dieron no solamente la vida sino también la capacidad de disfrutarla y encontrar en los pequeños logros, grandes satisfacciones y alegrías. Les agradezco todo su apoyo y comprensión, con su ejemplo me han enseñado a esforzarme para alcanzar mis metas, a vivir de manera honesta y a ver más allá de mis necesidades para ayudar a otros. Los amo con todo mi corazón y son mi razón de ser.

A mis hermanos, Alfonso, Juan Carlos y Alonso. Son mis mejores amigos y con quienes he compartido los buenos y malos momentos, son el apoyo constante por el cual he logrado llegar hasta aquí. Alf eres mi ejemplo a seguir no solamente como profesionista sino también como hermano; Juan Carlos, tu cariño le da alegría a mi vida y espero nunca pierdas el entusiasmo con el que haces las cosas que te gustan; Alonso, aunque aun eres pequeño sé que llegarás a ser un buen hombre por el espíritu alegre y fraternal que tienes.

A mi familia, quien conforma la mejor escuela en esta vida y de la que estoy orgulloso de ser parte. Les agradezco el amor y esfuerzo que dedican para que salga adelante, espero algún día poder corresponderles de la misma manera.

Al Dr. Manuel Martín Ortiz, mi asesor de tesis, que no solamente me impartió conocimientos sino que me contagió de su pasión por las Ciencias de la Computación. Sus enseñanzas y sabios consejos me han mantenido firme en este camino, ha sido una fuente de inspiración para mí y un buen amigo.

A Marco Ángel Vela Garay, mi amigo y compañero en este difícil camino, de quien he aprendido valiosas lecciones y ha compartido conmigo los grandes momentos de mi vida. Eres un gran ejemplo de esfuerzo y fortaleza para mí, y quiero compartir mis logros y metas contigo.

A Minerva Aidee Díaz Romero, mi mejor amiga, quien a pesar del tiempo y la distancia representa para mí una luz que me ayuda a ver en los momentos más oscuros. Te agradezco tu amistad y cariño incondicionales.

A mis amigos, que aunque no puedo nombrar a todos, son para mí uno de los mejores regalos de esta vida. Espero que sigamos unidos y podamos compartir nuestros logros.

A todos mis profesores, a quienes debo mi respeto y admiración. Han hecho de mí un profesionista apasionado por su carrera.

Introducción

La animación y modelado facial se refieren a técnicas para representar gráficamente el rostro humano en un sistema computacional y animar tal rostro de manera consistente con los humanos reales. Ésta es a menudo considerada una de las tareas más retadoras en el campo de la animación, ya que somos muy hábiles en identificar movimientos faciales innaturales, la más ligera inconsistencia provoca que la animación pierda su realismo.

En este trabajo se desarrolla un sistema que permite animar un rostro tridimensional, a partir de un conjunto inicial de modelos 3D que contiene distintas expresiones faciales y distintos visemas. El término visema viene del inglés *viseme* y fue acuñado como una amalgama de las palabras *visual* y *phoneme*, es decir, fonema visual. El sistema recibe un texto en español y genera la animación del rostro de tal manera que se logre la apariencia de que está pronunciando las palabras del texto. Para ello, el texto pasa por un proceso de silabeo, las sílabas se descomponen en fonemas y éstos son proyectados a su conjunto de visemas correspondiente para producir la animación.

El sistema está pensado para aplicarse en el campo de la educación especial, en particular en la educación de niños con hipoacusia. La hipoacusia es la disminución del nivel de audición por debajo de lo normal y con frecuencia da lugar a situaciones de minusvalía con importantes repercusiones físicas y psicológicas. En las instituciones públicas la carencia de infraestructura para la atención de estas dificultades es preocupante y existe la necesidad de atención para evitar el deterioro comunicativo relacional en su vida futura, por lo cual, el objetivo es dejar en claro la aplicación de la tecnología en estas áreas.

El trabajo se organiza de la siguiente manera: en el Capítulo 1 se presenta, a manera de introducción, una breve historia de la animación, relatando su evolución de la animación

convencional a la animación por computadora. Se presentan también, algunos principios de la animación que aun se aplican en la animación por computadora y se discuten los elementos más importantes de la animación 3D, empezando por el espacio de coordenadas tridimensional, las técnicas de modelado 3D y las distintas representaciones de los modelos, haciendo énfasis en la representación de mallas poligonales.

En el Capítulo 2, se estudia de manera particular la animación facial. Se presentan las características anatómicas del rostro para entender los factores que intervienen en la generación de expresiones faciales. Se discuten los distintos tipos de modelos faciales y las técnicas para crearlos. También, se exponen algunas de las técnicas de animación facial más utilizadas, en especial, la técnica de metamorfosis en 2D y 3D.

El Capítulo 3 está dedicado al estudio fonológico del idioma Español. Se describen los órganos del habla y la clasificación de los sonidos según las posibilidades de la laringe y el grado de constricción de la cavidad oral. Se presentan las principales generalizaciones del proceso de silabificación en el idioma Español, que sirven de base para el algoritmo de silabificación implementado. Por último, se discute la correspondencia entre fonemas y visemas, siendo éstos los elementos esenciales para la animación.

En el Capítulo 4 se establecen los objetivos generales y específicos del trabajo. Se describen además, el origen y la motivación detrás de la investigación. También, se discuten algunos trabajos previos en el campo de la educación especial realizados por diferentes investigadores en el estado de Puebla, para poner en contexto el presente trabajo. Al final del capítulo se exponen los requerimientos funcionales y no funcionales del sistema.

En el Capítulo 5 se presenta el proceso de diseño e implementación del sistema. Primero se discute el ambiente de desarrollo y las herramientas de trabajo. Después se describen las técnicas de metamorfosis y combinación de modelos 3D, la implementación del analizador fonológico del texto, el proceso de silabificación y la proyección de fonemas a visemas, las formas de asignar los tiempos de la animación, y la forma de añadir expresividad al rostro. Finalmente, se describe de manera general el funcionamiento del sistema.

El Capítulo 6 presenta los resultados obtenidos de las pruebas hechas con el sistema. Se evalúa el funcionamiento de cada uno de los procesos involucrados en la generación de la animación del rostro: el proceso de metamorfosis y combinación de modelos 3D, el proceso de

silabificación, la sustitución fonológica, la proyección a visemas y los tiempos de la animación. Al final se evalúa el desempeño del sistema como auxiliar en la terapia del lenguaje para niños con hipoacusia.

Posteriormente, se presentan las conclusiones del trabajo, así como sus limitaciones y perspectivas. Se incluyen dos apéndices, el Apéndice A es un manual de instalación y uso del sistema, mientras que el Apéndice B presenta una breve introducción a la interfaz de programación de aplicaciones (API) Java 3D.

Capítulo 1

Animación en 3D

La animación por computadora puede ser definida como una técnica en la cual la ilusión de movimiento es creada al mostrar en una pantalla o grabar en un dispositivo una serie de estados individuales de una escena dinámica [2]. En ella, cualquier valor que puede ser cambiado puede ser animado. La posición y orientación de un objeto son candidatos obvios para la animación, pero cualquiera de los siguientes puede ser animado también: la forma del objeto, los parámetros de sombreado, texturas, los parámetros de las fuentes de luz y los parámetros de la cámara [1].

Para poner la animación por computadora en contexto, es importante entender su herencia, su historia y ciertos conceptos relevantes. En la primera parte de este capítulo se discute la percepción de movimiento, la evolución técnica de la animación y sus principios. Al final se presentan conceptos del sistema tridimensional esenciales para entender la animación en 3D.

1.1. Percepción de movimiento

Cuando una serie de imágenes secuenciales separadas son puestas de manera secuencial y momentánea en frente de nuestros ojos parece que se mezclan y ya no las percibimos como varias imágenes sino como una imagen en movimiento. Las imágenes pueden comunicar una gran cantidad de información debido a que el sistema visual del ser humano es un procesador de información sofisticado. Por lo tanto, se sigue que las imágenes en movimiento tienen el potencial de comunicar aun más información en un tiempo corto. En efecto, el sistema visual del ser humano ha evolucionado para proveer supervivencia en un mundo siempre cambiante;

está diseñado para notar e interpretar el movimiento.

Cuando una animación es creada se graba típicamente en película o video como una serie de imágenes estáticas que cuando son mostradas en una secuencia rápidamente son percibidas por el observador como una sola imagen que se mueve. Esto es posible debido a que el complejo ojo-cerebro ensambla la secuencia de imágenes estáticas y la interpreta como un movimiento continuo. Una sola imagen presentada al observador por un corto tiempo dejará una impresión de ella misma -la imagen diferida positiva[1]- en el sistema visual por un período corto después de ser removida. Este fenómeno es atribuido a lo que se ha llamado *persistencia de visión*. Las retinas de nuestros ojos retienen la imagen de lo que vemos por una fracción de segundo después de que la imagen ha sido apartada de nuestra vista [3].

Cuando se le presenta a una persona una secuencia de imágenes estáticas estrechamente vinculadas a una velocidad suficientemente rápida, la persistencia de visión induce la sensación de imágenes continuas. Las imágenes diferidas de las tomas individuales llenan los vacíos entre las imágenes.

Tanto en película como en video, se graba una secuencia de imágenes para reproducirla a una velocidad lo suficientemente rápida para engañar al ojo y que las interprete como continuas. Cuando la percepción de imágenes continuas no se logra, se dice que la imagen *parpadea*[2]. Los receptores en el ojo continuamente muestrean la luz en el ambiente. Las limitaciones en la percepción del movimiento son determinadas por el tiempo de reacción de esos sensores y por otras limitaciones mecánicas tales como el parpadeo y el seguimiento. Si un objeto se mueve muy rápido con respecto al observador, entonces los receptores en el ojo no serán capaces de responder lo bastante rápido para que el cerebro distinga los detalles y resulta lo que se conoce como *falta de definición de movimiento*[3]. En secuencias de imágenes estáticas, la falta de definición de movimiento resulta como una combinación de la velocidad del objeto y el intervalo de tiempo sobre el cual la escena es muestreada. En una cámara estática, a un objeto que se mueve rápidamente no le faltará definición si la velocidad del obturador es lo bastante

[1] A single image presented to a viewer for a short time will leave an imprint of itself –the *positive afterimage*– in the visual system for a short time after it is removed. (Rick Parent, 2002).

[2] *When the perception of continuous imagery fails to be created, the image is said to flicker. (Rick Parent, 2002).*

[3] If an object moves too quickly with respect to the viewer, then the receptors in the eye will not be able to respond fast enough for the brain to distinguish sharply defined, individual detail; *motion blur* results. (Rick Parent, 2002).

rápida en relación con la velocidad del objeto. En los gráficos por computadora, la falta de definición de movimiento no se dará si es muestreada en un instante preciso en el tiempo.

En la animación se deben tomar en cuenta dos velocidades. Una es la velocidad de reproducción, que es el número de imágenes por segundo que se muestran en el proceso. La otra es la velocidad de muestreo, el número de imágenes diferentes que ocurren por segundo. Por ejemplo, una señal de televisión conforme al formato del National Television Standards Committee (NTSC) muestra las imágenes a una velocidad de treinta cuadros por segundo, pero en algunos programas puede que haya solamente seis imágenes diferentes por segundo, con cada imagen mostrada repetidamente cinco veces. Frecuentemente, la animación del movimiento de labios se hace cada dos cuadros ya que hacerla en cada cuadro hace que se vea muy agitada. Algunas películas muestran cada cuadro dos veces para reducir los efectos de parpadeo. Por otro lado, debido a que una señal de televisión NTSC es entrelazada (lo que significa que las líneas de trazado impares son mostradas empezando con el primer sexto de segundo y las líneas pares son mostradas empezando en el siguiente sexto de segundo), un movimiento más suave puede ser producido al muestrear la escena cada sexto de segundo aunque los cuadros completos son reproducidos únicamente a treinta cuadros por segundo [1].

Usar los principios de la persistencia de visión para crear la ilusión de movimiento fue un paso para entender el proceso de hacer películas. Las otras dos áreas principales de entendimiento y tecnología que necesitaban unirse a esto fueron la creación de un mecanismo para correr una secuencia de imágenes y otro para mostrar esta secuencia.

1.2. Una Breve Historia de la Animación

Para muchos en la industria del entretenimiento y los gráficos por computadora, la fascinación está en contar historias a través de imágenes en movimiento. Sin embargo, la creación de un movimiento convincente no es un problema trivial. Después de muchos años de progresos técnicos continuos, la animación es todavía un proceso de trabajo intensivo que requiere grandes habilidades y arte.

1.2.1. Los Primeros Dispositivos.

Uno de los más importantes componentes en el descubrimiento de la cinematografía fue la fotografía. Al igual que casi todos los otros descubrimientos a través del tiempo, la fotografía fue el resultado de una acumulación de conocimientos técnicos abarcando un periodo de no menos que trescientos años. De hecho, tal como el efecto de imagen a través de un agujero precedió la construcción de la cámara oscura, así también el conocimiento de sustancias sensibles a la luz precedieron la capacidad de capturar una imagen fija por medio de la fotografía. El efecto de la luz sobre compuestos de plata ha sido conocido desde 1727 cuando fue descubierto que los haluros de plata se tornaban negros cuando eran expuestos al sol. Pero la fotografía por sí sola tendría que esperar pacientemente hasta la llegada del celuloide antes de que el movimiento se pudiera conseguir.

Joseph Nicephore Niépce es acreditado por producir la primera "imagen" permanente del mundo, la cual llamó un heliógrafo (o dibujo de sol). La fotografía de Niépce fue hecha en 1826 y fue tomada desde una ventana con vista a los techos de las casas vecinas a la de Niépce (Figura 1-1). Él usó una placa de peltre que fue sensibilizada con betún de Judea. La fotografía fue hecha en una cámara oscura y tomó ocho horas de exposición. La fotografía reside en el Harry Ransom Humanities Center, The University of Texas en Austin. Fue descubierta por suerte en los años 50 en Londres junto a cartas escritas por Niépce [4].

Figura 1-1: Esta fotografía es parte de la Gernsheim Collection y es conocida como "View From The Window At Gras".

La persistencia de visión y la habilidad para interpretar una serie de imágenes como una imagen en movimiento fueron investigadas activamente en el siglo XIX, mucho antes de que la cámara de video fuera inventada. El reconocimiento y la subsecuente investigación de este efecto

llevó a una serie de dispositivos pensados como juguetes de sala. Tal vez el más simple de estos primeros dispositivos es el *thaumatrope* (su nombre en inglés), un disco plano con imágenes dibujadas en ambos lados y que tiene dos hilos conectados en el borde del disco (véase la Figura 1-2). El disco podía ser girado rápidamente tirando de los hilos. Cuando era girado lo bastante rápido, las dos imágenes parecían estar sobrepuestas. El ejemplo clásico usa la imagen de un pájaro en un lado y la imágen de una jaula en el otro; al girar el disco el pájaro se coloca visualmente dentro de la jaula. Una técnica igualmente primitva es el *flipbook*, una tableta de papel con un dibujo individual en cada página. Cuando las páginas son pasadas rápidamente, el observador tiene la impresión de movimiento.

Figura 1-2: Thaumatrope. De un lado se encuentra un pájaro y del otro una jaula, al girarlo rápidamente el pájaro se ve dentro de la jaula.

Uno de los primeros dispositivos de animación más conocidos es el *zoetrope*, *zoótropo* o rueda de la vida. El zoetrope tiene un cilindro pequeño y ancho que rota sobre su eje de simetría. Alrededor del interior del cilindro está una secuencia de dibujos, cada uno ligeramente diferente a los que están junto a él. El cilindro tiene largas hendiduras verticales entre cada par de imágenes adyacentes de tal forma que cuando es girado sobre su eje cada hendidura permite al ojo ver la imagen en la pared opuesta del cilindro (véase la Figura 1-3). La secuencia de hendiduras que pasan en frente del ojo mientras el cilindro es girado sobre su eje presenta una secuencia de imágenes al ojo, creando la ilusión de movimiento.

El descubrimiento de una forma para hacer películas usando un proceso fotográfico fue hecho por el inventor norteamericano, Thomas Edison, y su jóven asistente, William K. Laurie Dickson.

Dickson comenzó a trabajar en el proyecto bajo la dirección de Edison en 1888. Ellos lograron hacer una cámara de video que era capaz de tomar imágenes fijas, que llamaron Quinetoscopio.

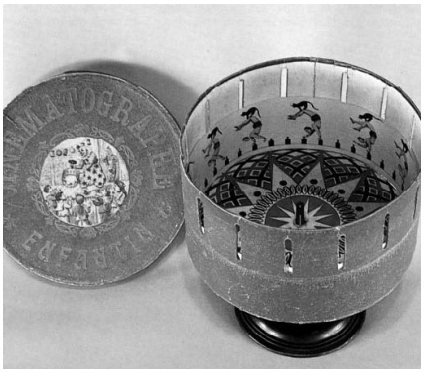

Figura 1-3: Un Zoetrope

Nadie en el campo de creación de imágenes en movimiento hubiera podido anticipar el impacto que la industria cinematográfica tendría en el próximo siglo. Sin embargo, Edison estaba consciente que el próximo paso lógico después de la cámara de video era crear un medio mecánico de proyección. Pero Edison le pidió a Dickson que hiciera a un lado el trabajo en este proyecto en favor de lo que el pensó que les dejaría más dinero en ese tiempo, su Quinetoscopio. Dickson después expresó su arrepentimiento ya que dejó el campo abierto y los hermanos Lumière pronto dieron el paso con su propio sistema de proyección de películas [3].

1.2.2. La Animación Convencional

La animación en Estados Unidos estalló en la forma de grabaciones de imágenes de dos dimensiones dibujadas a mano. La primera vez que se usó la cámara para hacer que cosas sin vida parecieran moverse ocurrió en 1896. George Méliès usó trucos simples de cámara tales como múltiples exposiciones y técnicas de paro de movimiento (*stop-motion techniques*) para hacer aparecer y desaparecer objetos y cambiar su forma. Algunos de los primeros pioneros de la animación cinematográfica fueron Emile Cohl, un francés que produjo varias ilustraciones; J. Stuart Blackton, un norteamericano que animó humo en una escena en 1900 (efectos especiales) y se le acredita la creación de la primera caricatura animada, en 1906; y el norteamericano Winsor McCay, el primer animador célebre, mejor conocido por sus trabajos de *Little Nemo* (1911) y *Gertie the Dinosaur* (1914). McCay es considerado por muchos como el productor de las primeras animaciones populares.

Como muchos de los primeros animadores, McCay fue un dibujante de caricaturas de periódico exitoso. Fue el primero en experimentar con color en la animación. Muchos de sus primeros trabajos fueron incorporados en actos de vodevil[4] en los que él interaccionaba con el personaje animado en la pantalla. De la misma manera, las primeras caricaturas frecuentemente incorporaban acción en vivo con personajes animados.

Los primeros avances técnicos importantes en el proceso de la animación pueden ser rastreados a los esfuerzos de John Bray, uno de los primeros en reconocer que patentar aspectos del proceso de animación resultaría en una ventaja competitiva. Empezando en 1970, su trabajo estableció las bases para la animación convencional tal como existe en la actualidad. Earl Hurd, quien juntó fuerzas con Bray en 1914, patentó el uso de cels[5] translúcidos en la composición de múltiples capas de dibujos en una imagen final y también patentó los dibujos en escala de grises opuestos a los de blanco y negro. Desarrollos posteriores de Bray y otros mejoraron la idea de incluir un sistema de clavijas para el registro y la de dibujar un fondo en largas hojas de papel para poder trasladar la cámara de forma paralela al plano del fondo (*panning*) con más facilidad.

Mientras la tecnología estaba avanzando, la animación como un arte seguía luchando. El primer personaje animado con una personalidad identificable fue Félix el Gato, dibujado por Otto Messmer de los estudios de Pat Sullivan. Sin embargo, a finales de los años 20, aparecen nuevas fuerzas: el sonido y Walt Disney.

Walt Disney fue, por supuesto, la fuerza dominante en la historia de la animación convencional. Su estudio no sólo contribuyó con varias innovaciones técnicas, sino que Disney, más que cualquier otro, avanzó la animación como un arte. Las innovaciones de Disney en la tecnología de la animación incluyen el uso de un *storyboard* para revisar la historia y bosquejos a lápiz para revisar el movimiento. Además, fue pionero en usar sonido y color en la animación (aunque no fue el primero en usar color). Disney también estudió secuencias de acción en vivo para crear movimiento más realista en sus películas. Cuando utilizó sonido por primera vez en *Steamboat Willie* (1928), ganó una ventaja sobre sus competidores.

[4] Comedia frívola, ligera y picante, de argumento basado en la intriga y el equívoco, que puede incluir números músicales y de variedades.

[5] Cel es abreviación de celuloide, que fue el material original usado para fabricar capas translúcidas. Actualmente, los cels están hechos de acetato.

Una de las innovaciones técnicas más significativas del estudio Disney fue el desarrollo de una cámara multi-plano. La cámara multi-plano consiste de una cámara montada arriba de múltiples planos, cada uno de los cuales contiene una celda de animación. Cada uno de los planos puede moverse en seis direcciones (derecha, izquierda, arriba, abajo, afuera, adentro), y la cámara puede moverse más cerca o más lejos.

La animación con la cámara multi-plano es más poderosa de lo que se podría pensar. Al mover la cámara más cerca de los planos mientras que los planos se utilizan para mover imágenes en primer plano hacia los lados, una amplificación más efectiva puede realizarse. Al mover múltiples planos a velocidades diferentes se puede producir el *efecto de paralaje*, que es el efecto visual de que los objetos más cercanos se mueven aparentemente más rápido a través del campo de visión que los objetos más alejados a medida que la mirada del observador se mueve de manera paralela a través del ambiente. Esto es muy efectivo para crear la ilusión de profundidad y una sensación mejorada de tres dimensiones. Al dejar abierto el lente de la cámara durante el movimiento se pueden producir varios efectos adicionales: las figuras pueden ser resaltadas a formas de una dimensión más alta; se pueden incorporar señales de profundidad a una imagen al reducir la nitidez de figuras en celdas más distantes; y se puede producir el efecto de falta de definición de movimiento (*motion blur*).

Con respecto al arte de la animación, Disney perfeccionó la habilidad para impartir personalidades únicas y atrayentes a sus personajes, como Mickey Mouse, Pluto, Goofy, y los siete enanos. Promovió la idea de que la mente del personaje era la fuerza impulsora de la acción y que un aspecto clave para el movimiento animado creíble era el análisis del movimiento de la vida real. También desarrolló piezas de humor, por ejemplo, *Skeleton Dance* (1929) y *Fantasia* (1940).

La rica herencia de la animación dibujada a mano en los Estados Unidos hace que naturalmente se le considere como la precursora de la animación por computadora, la cual también tiene fuertes raices en este país. Sin embargo, la animación por computadora también tiene una estrecha relación con otras técnicas de animación. Una comparación cercana puede hacerse entre la animación por computadora y algunas técnicas de paro de movimiento (*stop-motion*), tales como animación con títeres y arcilla. De manera típica en la animación por computadora, el primer paso es el proceso de modelado del objeto. Los modelos son después manipulados

para crear las escenas tridimensionales que son procesadas para producir las imágenes de la animación.

De forma muy parecida, la animación con arcilla y títeres con paro de movimiento usan figuras tridimensionales que son construidas y luego animadas en escenarios separados y bien definidos. Una vez que las figuras físicas tridimensionales son creadas, son usadas para presentar un ambiente tridimensional. Una cámara se posiciona para ver el ambiente y grabar una imagen. Una o más figuras son manipuladas, y la cámara se puede reposicionar. La cámara graba otra imagen de la escena. Las figuras son manipuladas de nuevo, y otra imagen se toma de la escena, y el proceso se repite para producir la secuencia de animación. Willis O'Brien de *King Kong* es considerado generalmente como el decano de este tipo de animación con paro de movimiento. Ejemplos impresionantes más recientes de la animación en 3D con paro de movimiento son la serie *Wallace and Gromit* de Nick Park y las producciones de Tim Burton *Nightmare Before Christmas* y *James and the Giant Peach* [1].

1.2.3. La Animación por Computadora

Tal vez uno de los primeros pioneros de la animación por computadora fue Lee Harrison III. A principios de los años 60, experimentó con figuras animadas usando circuitos analógicos y un tubo de rayos catódicos. Adelantado a su tiempo, él armó un traje con potenciómetros y creó el primer aparejo para captura de movimiento funcional, animando figuras 3D en tiempo real en su pantalla CRT[6]. Él realizó varios cortometrajes con este sistema, llamado ANIMAC. Éste evolucionó a SCANIMATE el cual comercializó con gran éxito en 1969, SCANIMATE permitía un control interactivo (escalamiento, rotación, traslación), grabar y reproducir elementos de video en capas para generar animaciones 2D y logotipos voladores para la televisión. La mayoría de los logotipos voladores 2D y de elementos gráficos para la publicidad televisiva de los años 70 fueron producidos usando los sistemas SCANIMATE.

El siguiente sistema extenso fue GRASS (*Graphics Symbiosis System*) desarrollado por Tom DeFanti para su tesis doctoral de 1974 en la Ohio State University. GRASS era un lenguaje para especificar animación 2D de objetos y aunque no era interactivo, era el primer sistema

[6] Abreviación de *cathode-ray tube* (tubo de rayos catódicos). Un CRT funciona moviendo un rayo de electrones a través de la pantalla. Cada vez que el rayo pasa sobre la pantalla, enciende puntos de fósforo en el interior del tubo de vidrio, iluminando las porciones activas de la pantalla.

gratuito que podía ser dominado por un usuario sin conocimientos técnicos. Con GRASS, se podía hacer un *script* para escalar, trasladar, rotar y hacer cambios de color de objetos 2D a tráves del tiempo. Pronto se convirtió en un gran éxito entre la comunidad artística que estaba experimentando con el nuevo medio de CG[7]. En 1978 se actualizó para trabajar en 3D con áreas sólidas y volúmenes y corría en una computadora de casa Bally. Esta versión se llamó ZGRASS, y también fue importante al traer los gráficos y la animación por computadora a la comunidad artística en plataformas computacionales comprables. La tecnología usada en las secuencias de la primera película de *Star Wars* (1977) se derivó de GRASS.

También en 1974, Nestor Burtnyk y Marcelli Wein desarrollaron un sistema de animación por computadora experimental en la National Film Board of Canada, permitía a los artistas animar dibujos de líneas 2D capturados en una tabla de datos. La animación se hacía interpolando punto a punto líneas correspondientes en una serie de fotogramas clave. El sistema fue usado para el clásico cortometraje de 1974 *Hunger*, que fue la primer historia animada usando animación por computadora nominada a los premios de la Academia.

El New York Institute of Technology Computer Graphics Lab (NYIT), entonces bajo la dirección de Ed Catmull, expandió esta idea, produciendo un sistema de animación comercial llamado TWEEN. Como el sistema de la National Film Board, TWEEN era un sistema 2D que permitía al animador dibujar fotogramas clave, y la computadora interpolaba los segmentos de línea correspondientes entre ellos. TWEEN automatizó el proceso de producir fotogramas intermedios, pero aún requería el talento de un artista o animador entrenado para los fotogramas clave. Aunque este método aceleró el proceso de animación a mano, las animaciones producidas de esta forma tenían un aspecto fluido excesivamente distintivo y el método no fue extensamente adoptado para la animación comercial [5].

Los primeros sistemas de animación 3D completos pueden ser categorizados en dos tipos de sistemas: sistemas programados y sistemas interactivos de fotogramas. El primer tipo fue ejemplificado por ANIMA-II [6] y ASAS [7]. Ambos usaban un lenguaje de programación para describir una secuencia de eventos y funciones en el tiempo. Al ser evaluadas sobre el tiempo y hacer una toma de la animación, producían la animación deseada.Vale la pena señalar que muchas de las secuencias de CG en la película de 1982 de Disney *TRON* fueron animadas con

[7] Abreviación de *Computer Graphics*.

ASAS. Estos sistemas eran poderosos ya que casi todo podía hacerse si podía ser programado, pero limitados ya que se requerían habilidades de programación para dominarlos.

Los sistemas de fotogramas clave eran más favorables para los artistas de la animación. Basados en la aproximación de fotogramas clave de la animación tradicional, estos sistemas permitían al usuario posicionar objetos y figuras en la escena de forma interactiva, salvar estas posiciones como fotogramas clave y la computadora calculaba los fotogramas intermedios para producir la animación final. GRAMPS[8] y BBOP[9] fueron ejemplos de este tipo de sistema. Ambos contaban con la interactividad del Evans & Sutherland Multi-Picture System, un excelente sistema de visualización de gráficos vectoriales que trabajaba a partir de una lista de presentación permitiendo actualizaciones instantáneas de los gráficos en la pantalla.

GRAMPS fue desarrollado para visualizar estructuras químicas aunque O´Donnell da ejemplos de cómo puede ser usado para animar una figura humana. Aunque claramente un sistema de script interpretado, GRAMPS permitía que variables del script estuvieran conectadas a botones para manipularlas de manera interactiva.

BBOP fue desarrollado en el NYIT por Garland Stern expresamente para la animación de personajes y fue usado extensamente por NYIT en producciones comerciales por 6 años. En BBOP, los animadores podían controlar de forma interactiva las transformaciones de las articulaciones en una jerarquía 3D, salvando las poses en fotogramas clave los cuales la computadora podía interpolar para producir una animación suave. El sistema respondía bien y era fácil de usar, y animadores convencionales entrenados produjeron varias buenas animaciones notables que aparecieron en varias películas en SIGGRAPH[8].

La mayoría de los sistemas de fotogramas comerciales modernos están basados en la aproximación de fotogramas interactivos de BBOP con características agregadas que facilitan el proceso de animación. En su núcleo, todos tienen características de BBOP (algunas copiadas, algunas desarrolladas independientemente), incluyendo las estructuras de esqueleto jerárquicas, la actualización interactiva en tiempo real de valores de transformación, la interpolación de fotogramas por canales de tal forma que diferentes articulaciones pueden tener diferentes claves en diferentes fotogramas, la elección de funciones de interpolación tales como lineales, cúbicas,

[8]SIGGRAPH es el Associaton for Computing Machinery´s (ACM´s) Special Interest Group on Computer Graphics. La ACM es el principal grupo profesional para científicos de la computación.

ease-in, ease-out, reproducción inmediata y un editor de interpolación.

En general, sin embargo, los sistemas por script son todavía mejores para los movimientos repetitivos o que son fáciles de describir, pero requieren habilidades de programación más allá de las habilidades de la mayoría de los artistas, especialmente a medida que los movimientos se van haciendo más complejos. Hacer un script para personajes expresivos, por ejemplo, es extremadamente difícil. Los sistemas interactivos de fotogramas son justamente lo opuesto. Éstos permiten a los artistas interaccionar directamente con los objetos y las figuras en un marco conceptual familiar. Pero se hacen ineficientes o tediosos para movimientos mecánicos o de algoritmos complejos. Debido a que son más usados por los artistas, la aproximación de fotogramas interactivos ha ganado el mercado de software comercial. Curiosamente, a medida que los animadores se vuelven más sofisticados en el uso de la animación por computadora, las capacidades para hacer scripts están empezando a reaparecer en los sistemas de fotogramas.

Los primeros sistemas de animación 3D trataban en su mayoría con cinemática hacia adelante simple (*forward kinematics*) de cuerpos articulados, sin embargo, la cinemática inversa (*inverse kinematics*) puede ser también un elemento importante en un equipo de animación. Al mover solamente una mano o un pie, el animador puede posicionar un miembro entero. Michael Girard construyó un sistema de animación sofisticado con cinemática inversa para su tesis doctoral[10] que fue usado para producir movimientos del cuerpo humano con gracia en su película de 1989 *Eurhythmy*. Posteriormente comercializó su sistema como un plug-in de 3D Studio MAX, Biped (parte del paquete Character Studio), donde la locomoción con piernas tal como caminar, correr y saltar puede ser animada al colocar huellas. Sus algoritmos de cinemática inversa calculan los movimientos de la figura que causan que siga las huellas.

La dinámica es también una herramienta importante para una animación realista. Jane Wilhelms fue una de las primeras en demostrar el uso de la dinámica para controlar un personaje animado [11]. Desde entonces, James K. Hahn[12], David Baraff[13] y Michael McKenna[14] han descrito una dinámica robusta para la animación por computadora. No obstante, apenas en los últimos años los principales sistemas comerciales han incorporado dinámica en su software. Los problemas que enfrentan son cómo integrar controles de dinámica, de cinemática inversa, y de cinemática hacia adelante en un mismo sistema, y presentar y resolver los conflictos potenciales de fuerzas que cada uno pone sobre los elementos animados.

La cinemática y la dinámica tratan con estructuras de esqueleto articuladas. Sin embargo, no toda la animación se hace con esqueletos. Una cara, por ejemplo, es una superficie única con deformaciones complejas. En el Capítulo 2 se discuten algunas técnicas de animación facial.

Una técnica de graficación popular para efectos especiales es el uso de sistemas de partículas. William T. Reeves[15] desarrolló un método donde usa un torrente controlado aleatorio de partículas para simular fuego, pasto, chispas, fluidos y una serie de otros fenómenos naturales. Fue usado en *Star Trek II: The Wrath of Khan* (1982, Lucasfilm), en la cual una pared de fuego barre la superficie de un planeta. Aunque para el estándar actual la pared no es muy convincente, fue un paso importante en el uso de gráficos por computadora en películas. Los sistemas de partículas son fáciles de implementar y aparecieron rápidamente en muchas animaciones de aficionados, académicos y profesionales de CG, especialmente en *Particle Dreams* de 1988 por Karl Sims. Los sistemas de animación comerciales tardaron un poco más en incorporar la técnica en sus estructuras establecidas, pero hoy en día todos la tienen en una forma u otra.

Otras técnicas de animación para efectos específicos en la literatura incluyen paso automático (caminar, correr, saltar, etc.) [17][16], comportamiento de bandadas [18], flujo de fluidos [19], olas [20], humo [21], arena [22], objetos flexibles [23], víboras [24], ropa [25] y muchos más.

La animación más difícil es la animación de personajes, particularmente la animación de personajes humanos. En la búsqueda de un movimiento más realista, la gente ha optado por grabar directamente los movimientos de un actor. Lee Harrison III en los años 60 fue el primero de muchos en usar este concepto. En 1983 Ginsberg y Maxwell presentaron un sistema de animación usando una serie de LED[9] pegados a un actor. Un conjunto de cámaras triangulaban las posiciones de los LED, obteniendo un conjunto de puntos 3D en tiempo real. El sistema fue usado poco tiempo después para animar a un presentador de television CG en Japón. Sin embargo, los sistemas de captura de movimiento y las computadoras gráficas no eran lo bastante rápidos entonces para las demandas de tiempo real de la animación actuada.

Cuando empezaron a volverse lo suficientemente rápidos, alrededor de 1988 con la presentación de las estaciones de trabajo Silicon Graphics 4D, tanto deGraf/Wharman y Pacific Data Images desarrollaron controles mecánicos para manejar personajes CG animados. Por varias razones la tecnología y el mercado no estaban listos y los sistemas fueron raramente

[9] LED, *Light-emitting diode*, diodo emisor de luz.

explotados después de su uso inicial.

Entonces, a inicios de los años 90, SimGraphics, Medialab y Brad deGraf, todos desarrollaron independientemente sistemas que permitían a actores en vivo controlar las acciones de un personaje CG en tiempo real. Estos sistemas permitían animar a los personajes en vivo, así como para procesamiento posterior. La animación puede generarse rápidamente por actores y titiriteros bajo el control de un director que tiene realimentación inmediata de la versión en tiempo real del personaje. Los tres sistemas han sobrevivido a sus versiones iniciales y aplicaciones y continúan siendo usados exitosamente en proyectos comerciales.

Al principio, estos sistemas existían por su cuenta y no estaban integrados en otros sistemas CG comerciales. La animación hecha en un sistema de fotogramas no podía ser mezclada fácilmente con animación realizada en un sistema de tiempo real. Con el paso del tiempo, tanto los sistemas de tiempo real como los sistemas de fotogramas han evolucionado, ahora muchos sistemas de fotogramas tienen provisiones para entrada en tiempo real y los sistemas de tiempo real importan y exportan curvas de animación por fotogramas.

Similar a la animación actuada, pero sin la realimentación en tiempo real, están los sistemas de captura de movimiento. Éstos son generalmente sistemas ópticos que usan marcadores reflectores en un actor. Durante la actuación, múltiples cámaras calculan las posiciones 3D de cada marcador, siguiéndolo a través del espacio y el tiempo. Un proceso fuera de línea empareja estos marcadores con posiciones en un esqueleto CG, duplicando el movimiento realizado. Aunque existen problemas al perder marcadores debido a las obstrucciones temporáneas y el proceso de emparejamiento puede ser muy laborioso, la captura de movimiento permite un procesamiento exacto del movimiento del cuerpo humano, particularmente cuando se trata de simular el movimiento de un actor en particular como lo que hizo Digital Domain con el video musical de Michael Jackson de 1997, *Ghosts*.

Con la llegada de computadoras de escritorio a bajo precio que procesan video, la animación está al alcance de más personas que nunca. Queda por ver cómo los límites de la tecnología serán empujados a medida que nuevas e interesantes formas de crear imágenes en movimiento son exploradas.

1.3. Principios de la Animación

Para estudiar algunas técnicas y algoritmos usados en la animación por computadora, es útil entender primero su relación con los principios de la animación usados en la animación convencional. Entre finales de los años 20 y los años 30 la animación pasó de una novedad a una forma de arte en los estudios de Walt Disney. Con cada película, las acciones se volvían más convincentes, y los personajes iban emergiendo como verdaderas personalidades. La audiencia era entusiasta y muchos de los animadores estaban satisfechos, sin embargo, era claro para Walt Disney que el nivel de animación y los personajes existentes no eran adecuados para perseguir nuevas historias, una nueva aproximación de dibujo era necesaria para mejorar el nivel de animación.

Disney estableció clases de dibujo para sus animadores en el Chouniard Art Institute en Los Angeles bajo la instrucción de Don Graham. Cuando iniciaron las clases, la mayoría de los animadores dibujaban usando la fórmula de las viejas caricaturas con figuras, tamaños, acciones, y gestos estandarizados, con poca o ninguna referencia a la naturaleza. De estas clases nació una forma de dibujar figuras humanas y animales en movimiento. Los alumnos estudiaban modelos de movimiento, así como también películas de acción en vivo, reproduciendo ciertas acciones una y otra vez. El análisis de la acción se volvió importante en el desarrollo de la animación.

Algunos de los animadores comenzaron a aplicar las lecciones de estas clases para animación de la producción, la cual se hizo más sofisticada y realista. Los animadores buscaban continuamente mejores formas para comunicarse entre sí las ideas aprendidas en estas lecciones. Gradualmente, los procedimientos se aislaron y fueron nombrados, analizados y perfeccionados, y se les enseñaban a los nuevos artistas. Se convirtieron en los principios fundamentales de la animación tradicional [26].

1.3.1. Simulación física

Un objeto dado posee algún grado de rigidez y debe aparentar tener alguna cantidad de masa. Esto es reflejado en la distorsión de su forma durante una acción, especialmente una colisión. La animación debe apoyar estas nociones consistentemente para un objeto dado a través de la animacón. Las siguientes técnicas establecen la base física para los objetos en una

escena.

Aplastar y Ensanchar (Squash & Stretch)

Cuando un objeto se mueve, el movimiento enfatiza cualquier rigidez del objeto. En la vida real, sólo las figuras más rígidas (como sillas o platos) permanecen así durante el movimiento. Cualquier cosa compuesta de tejido vivo, sin importar cuántos huesos tenga, mostrará un movimiento considerable en su forma durante una acción. Por ejemplo, un rostro, ya sea masticando, sonriendo, hablando, o simplemente mostrando un cambio de expresión, presenta varios cambios de figuras en las mejillas, los labios, y los ojos.

La posición aplastada representa la forma ya sea aplanada por una presión externa o estrechada por su propia fuerza. La posición ensanchada siempre muestra la misma forma en un estado muy extendido. La regla más importante para aplastar y ensanchar es que, sin importar qué tan aplastado o ensanchado se vuelve un objeto en particular, su volumen permanece constante. Si un objeto se aplasta sin ensancharlo de los lados, parecerá que se encoge; si es ensanchado hacia arriba sin aplastarlo de los lados parecerá que crece. Una prueba estándar en la animación para los principiantes es dibujar una pelota rebotando como se muestra en la Figura 1-4.

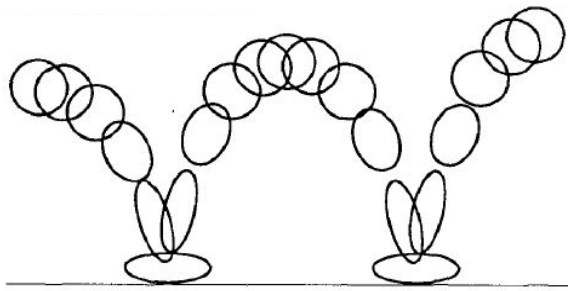

Figura 1-4: Aplastar y ensanchar para animar una pelota rebotando. Si el dibujo inferior es aplastado, da la apariencia de que rebota.

Aplastar y ensanchar también define la rigidez del material del cual está compuesto un objeto. Cuando un objeto es aplastado y ensanchado drásticamente, da la apariencia de que el objeto está compuesto de un material suave y flexible. Cuando las partes de un objeto son de diferentes materiales, deben responder de diferente forma: las partes flexibles deben aplastarse

más que las partes rígidas.

En la animación facial el aplastar y ensanchar es muy importante, no únicamente por mostrar la flexibilidad de la piel y los músculos, sino también por mostrar la relación entre las partes del rostro. Cuando el rostro sonríe anchamente, las esquinas de la boca empujan las mejillas. Las mejillas se aplastan y empujan hacia arriba en dirección a los ojos, produciento un guiño, lo que hace que las cejas bajen y la frente se ensanche. Cuando el rostro adopta una expresión de sorpresa, la boca se abre, ensanchando las mejillas. Los ojos bien abiertos empujan las cejas hacia arriba, aplastando y arrugando la frente.

Medida del Tiempo (Timing)

La medida del tiempo, o la velocidad de una acción, es un principio importante porque le da significado al movimiento, la velocidad de una acción define qué tan bien la idea detrás de una acción será entendida por la audiencia. Refleja el peso y tamaño de un objeto, e incluso puede transmitir un significado emocional.

Más que cualquier otro principio, la medida del tiempo define el peso de un objeto. Dos objetos, idénticos en tamaño y forma, pueden parecer ser de distintos pesos tan sólo con manipular el tiempo. Entre más pesado es el objeto, más grande su masa, y más fueza se requiere para cambiar su movimiento. Un cuerpo pesado es más lento para acelerar y desacelerar que uno liviano. Cuando se trata de objetos pesados, uno debe permitr bastante tiempo y fuerza para empezar, detener o cambiar sus movimientos, para lograr que su peso luzca convincente.

La forma en que un objeto se comporta en la pantalla, el efecto de peso que da, depende totalmente en el espaciado de las poses y no en las poses mismas. No importa qué tan bien esté dibujada una bala de cañón, no parecerá una bala si no se comporta como tal en la animación. Lo mismo aplica para cualquier objeto o personaje.

Acción secundaria

Una acción secundaria es una acción que resulta directamente de otra acción. Las acciones secundarias son importantes en realzar el interés y añadir una compejidad más realista a la animación. La expresión facial de un personaje será en ocasiones una acción secundaria. Cuando la idea principal de una acción se transmite a tráves del movimiento del cuerpo, la expresión

facial pasa a ser subordinada a la idea principal. Si esta expresión se va a animar o cambiar, el peligro no está en que la expresión domine la escena, sino en que nunca sea vista. El cambio debe venir antes, o después del movimiento. Un cambio a la mitad de un movimiento más importante pasará desapercibido. Las acciones secundarias sirven de apoyo a la acción principal, posiblemente proporcionando reacciones físicas a la acción anterior.

Lentificar hacia adentro y hacia afuera (Slow In and Out)

Lentificar hacia adentro y hacia afuera se ocupa del espaciado de los dibujos intermedios entre las poses en los extemos. Matemáticamente, el término se refiere a continuidad de movimiento de segundo y tercer orden.

En la animación temprana, la acción era limitada a movimientos rápidos y lentos principalmente, el espaciado de un dibujo al siguiente era bastante parejo. Pero cuando las poses se volvieron más expresivas, los animadores quisieron que la audiencia las viera. Descubrieron que al agrupar los dibujos intermedios podían lograr un resultado bastante enérgico. Lentificar hacia afuera una pose, después lentificar hacia adentro la siguiente simplemente se refiere al tiempo entre los dibujos intermedios.

El animador señala la colocación de los intermedios, la lentitud hacia afuera y hacia adentro, con una tabla de tiempos dibujada al costado del dibujo como en la Figura 1-5.

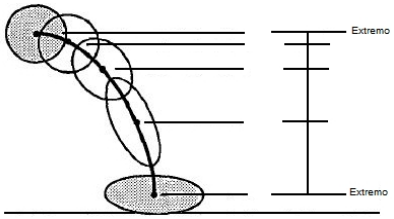

Figura 1-5: Señalamiento de tiempo para una pelota rebotando.

En la mayoría de los sistemas de animación por computadora 3D con fotogramas, el espaciado intermedio se realiza automáticamente usando interpolación con splines. Lentificar hacia afuera y adentro se logra al ajustar la tensión, dirección o sesgo y continuidad de los splines. Esto funciona bien para dar el efecto de lentitud hacia afuera y adentro, pero se requiere una representación gráfica del spline para ver el efecto que la tensión, dirección y continuidad tendrán

en su forma.

Arcos

La trayectoria visual de la acción de un extremo a otro siempre es descrita por un arco. En la naturaleza los arcos son las rutas más económicas por las cuales una forma puede moverse de una posición a otra. Los objetos debido a las leyes de la Física tales como la gravedad, usualmente se mueven en arcos en lugar de líneas rectas. En la animación, los arcos son usados extensamente, ya que hacen que la animación sea más suave y menos rígida que una línea recta para la trayectoria de la acción.

1.3.2. Diseño de Acciones Estéticas

Frecuentemente el animador necesita exagerar un movimiento para que no pase desapercibido o para transmitir una idea. Para mantener la atención del público, el animador necesita hacer que sea agradable a la vista. Además, las acciones deben fluir de una a otra para hacer que la toma entera parezca evolucionar continuamente en lugar de verse como movimientos separados.

Exageración

El significado de exageración es, en general, obvio. Sin embargo, el principio de exageración en la animación no significa distorsionar formas y objetos de manera arbitraria o hacer una acción violenta o irreal. El animador debe ir al corazón de cualquier cosa o cualquier idea y desarrollar su esencia, entendiendo su razón de ser, para que la audiencia también la entienda. Una escena tiene muchos componentes: el diseño, la forma de los objetos, la acción, la emoción, el color, el sonido. La exageración puede funcionar con cualquier componente, pero no de manera aislada. La exageración de los distintos componentes debe estar balanceada.

Atracción

La atracción significa cualquier cosa que una persona disfruta ver: una cualidad de encanto, diseño placentero, simplicidad, comunicación, o magnetismo. El ojo es atraído a una figura u objeto atractivo, y una vez ahí, se mantiene para apreciar el objeto. Un dibujo o diseño soso

carece de atracción. Un diseño demasiado complicado o difícil de entender también carece de atracción.

Al crear una pose atractiva para un personaje, una cosa que se debe evitar son los llamados "gemelos", donde ambos brazos y ambas piernas están en la misma posición, haciendo la misma cosa. Esto le da a la pose una cualidad rígida y poco atractiva. Si cada parte del cuerpo varía de alguna forma de su parte correspondiente, el personaje se verá más natural y más atractivo.

Acción de seguimiento y traslapo (Follow Through and Overlapping Action)

Las acciones raramente se detienen completamente de manera repentina, generalmente son llevadas más allá de su punto final. Por ejemplo, una mano, después de lanzar una pelota, continúa más allá del punto en donde se soltó la pelota.

En el movimiento de cualquier objeto o figura, las acciones de las partes no son simultáneas: algunas partes deben iniciar el movimiento, éstas conforman la delantera. Al caminar, la acción empieza con las caderas. Al moverse hacia adelante, hacen que las piernas se muevan. La cadera "guía", y las piernas "le siguen". Los apéndices o partes sueltas de un personaje u objeto se moverán más lento y arrastrarán detrás de la parte que guía de la figura. Después cuando la parte guía de la figura se detiene, estos apéndices continuarán moviéndose y tardarán más en detenerse.

El traslapo es crítico para comunicar ideas principales de la historia. Una acción nunca debe detenerse por completo antes de iniciar otra acción, y la segunda acción debe traslaparse con la primera. El traslapo mantiene un flujo continuo entre frases o acciones completas.

1.3.3. Presentación Eficaz de las Acciones

Anticipación

Una acción ocurre en tres partes: la preparación de la acción, la acción en sí, y su finalización. La anticipación es la preparación de la acción, y tiene muchas facetas. En un sentido, es la disposición anatómica para una acción. Debido a que los músculos del cuerpo funcionan a través de contracción, cada uno debe ser primero extendido antes de poder contraerse. Sin ninguna anticipación muchas acciones son abruptas, rígidas e innaturales.

Figura 1-6: Tomar un vaso. Note como el brazo se dobla antes de estirarlo y como la mano se abre en anticipación a tomar el vaso.

La anticipación también es un mecanismo para atrapar la atención del público, y prepararlo para el siguiente movimiento haciendo que lo esperen antes de que ocurra, así la atención de la audiencia se dirige a la parte correcta de la pantalla en el momento correcto (Figura 1-6).

También puede enfatizar el peso, como cuando un personaje recoge un objeto que es muy pesado. Una anticipación exagerada, como agacharse mucho para recoger el objeto, ayuda al momento del personaje para cargarlo.

Representación (Staging)

Es la presentación de una idea de forma que es completa e inequívocamente clara, este principio viene directamente de la animación a mano en 2D. Una acción es representada para ser entendida; una personalidad es representada para ser reconocida; una expresión para ser vista; un humor para afectar a la audiencia.

Para representar una idea clara, la atención de la audiencia debe ser dirigida exactamente a donde debe de estar en el momento correcto. La representación, anticipación y sincronización son integrales para dirigir el ojo. Al representar una acción, es importante que sólo una idea sea vista por el público en un momento dado. Si muchas acciones suceden al mismo tiempo, el ojo no sabe a dónde mirar y la idea principal de la acción es pasada por alto.

1.3.4. Técnicas de Producción

Existen dos aproximaciones principales en la animación a mano: la acción hacia adelante o todo derecho (*Straight Ahead Action*) y la acción pose-a-pose.

La acción hacia adelante se conoce así debido a que el animador literalmente trabaja hacia

adelante a partir de su primer dibujo en la escena. Él sabe dónde encaja la escena en la historia y lo que tiene que incluir. Hace un dibujo después de otro, dándose ideas mientras continúa, hasta que alcanza el final de la escena.

La segunda aproximación es llamada pose-a-pose. En ella el animador planea sus acciones, y establece qué dibujos serán necesarios para la animación, hace los dibujos concentrándose en las poses, los relaciona entre sí en tamaño y acción, y luego dibuja los intermedios. La aproximación pose-a-pose es usada en la animación que requiere una buena actuación, donde las poses y el tiempo son importantes.

La técnica pose-a-pose se aplica en la animación por computadora de fotogramas clave con controles para el tiempo y las poses de los extemos y los intermedios. La dificultad para controlar los intermedios hace que sea incorrecto aproximar la animación por computadora exactamente como uno haría con la animación a mano pose-a-pose. Cuando se trabaja con un modelo complejo, crear una pose completa en un momento hará que los intermedios sean demasiado impredecibles. La trayectoria de la acción será generalmente incorrecta y los objetos se intersecarán. El resultado es un trabajo de bastante tiempo para arreglar los intermedios.

Existe una mejor aproximación en el contexto de un sistema de modelado jerárquico, que trabaja capa a capa sobre la jerarquía. En lugar de animar una pose completa a otra, se anima una transformación a la vez, empezando por el tronco de la estructura de árbol jerárquica, trabajando transformación a transformación hacia abajo en las ramas hasta el final. Esta aproximación por capas comparte muchos elementos importantes con la técnica pose-a-pose en la animación a mano. Planificar la animación previamente, como en pose-a-pose, se torna incluso más importante. Se debe pensar bien la acción, planificar los tiempos y poses para que incluso en las primeras capas, las poses y acciones sean claras.

1.4. Animación por Fotogramas Clave (Keyframe Animation)

Para entender este término, se debe entender que el tiempo, en las películas, se mide en fotogramas. Hay 24 fotogramas por segundo (f.p.s. por su nombre en inglés *frames per second*) en las películas proyectadas en el cine, 30 f.p.s. en la televisión y video en el formato NTSC, y 25 f.p.s. en el formato de televisión y video PAL y frecuentemente, 15 f.p.s. o incluso menos

para animación en línea o películas.

El nombre de fotograma clave o *keyframe* viene de la animación a mano tradicional donde el animador principal (al ser el mejor artista) dibujaba las poses clave (o extremos de posición y movimiento) de un personaje para un número de fotograma en particular (un punto dado en el tiempo) para marcar cada cambio importante en la posición de un personaje. Otro artista entonces rellenaba las posiciones intermedias. Los fotogramas clave ocurren en la secuencia muy a menudo, así que la acción intermedia está razonablemente bien definida, o las claves son acompañadas por información adicional para indicar la colocación de los fotogramas intermedios. En la animación por computadora, el término fotograma clave ha sido generalizado para aplicarse a cualquier variable cuyo valor se establece en un fotograma clave específico y a partir de la cual los valores para los fotogramas intermedios son interpolados de acuerdo a algún procedimiento prescrito. Estas variables han sido referidas en la literatura como variables de articulación (*articulation variables* o *avars*) y los sistemas son a veces referidos como basados en pistas (*track based*) [1]. Es común que estos sistemas provean una interfaz interactiva con la cual el animador pueda especificar los fotogramas clave y la interpolación deseada.

1.5. El Sistema 3D

Podemos relacionarnos con un espacio tridimensional debido a que vemos nuestro mundo en 3D. En nuestro mundo, los objetos no tienen solamente anchura y altura, como lo hacen en el espacio 2D, también tienen una profundidad asociada y pueden localizarse cerca o lejos de nosotros.

Cada punto en un mundo 3D se localiza usando tres valores de coordenadas. Para poder definir puntos con tres coordenadas se define un tercer eje, llamado por lo regular eje Z. El eje Z es perpendicular a los ejes X y Y. Los tres ejes se encuentran en el origen definido como (0,0,0) como se muestra en la Figura 1-7.

Las escenas de CG representan objetos de muchos tipos: personas, árboles, flores, fuego, etc. Cada uno tiene características radicalmente diferentes, así que no debe sorprender que no haya una metodología estándar para el modelado, cada objeto requiere un tratamiento diferente para hacerle justicia al modelo.

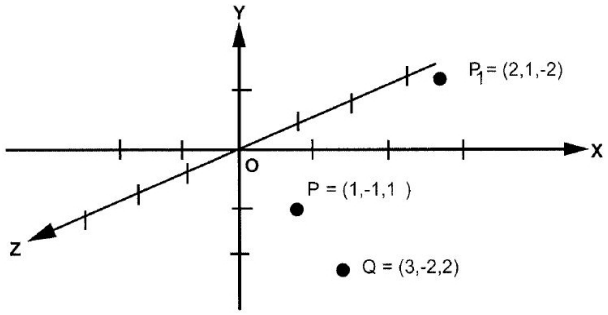

Figura 1-7: El sistema de coordenadas en tres dimensiones.

Los objetos 3D son más complejos de manejar que otra información multimedia, como señales de audio o imágenes 2D, ya que existen muchas representaciones distintas para tales objetos. Un objeto puede ser representado en una rejilla 3D como una imagen digital, o en el espacio Euclidiano 3D. En este último caso, el objeto puede ser expresado por una sola ecuación (como las superficies implícitas algebraicas), por un conjunto de facetas representando su superficie límite o por un conjunto de superficies matemáticas. La Figura 1-8 ilustra varias representaciones de un conejo en 3D.

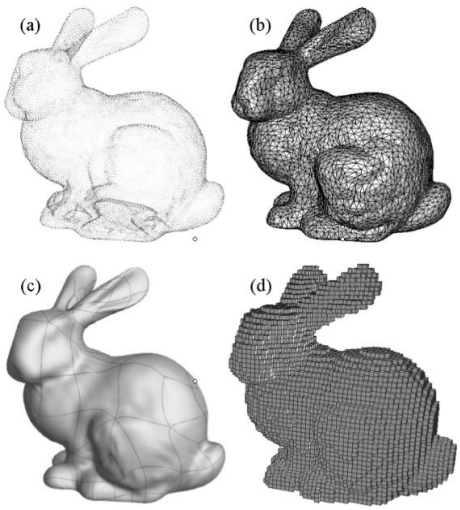

Figura 1-8: Diferentes representaciones para un conejo: (a) una nube de puntos, (b) una malla triangular, (c) un conjunto de superficies paramétricas, (d) un conjunto de voxels (amalgama de las palabras volumétrico y píxel).

La Figura 1-8(a) ilustra el uso de una nube de puntos para representar al conejo. Esta representación basada en puntos, provista por un escáner, no es eficiente para computar propiedades físicas o geométricas o para mostrarse en una pantalla. De hecho no hay relaciones de vecindad entre los puntos 3D, ni tampoco alguna información de superficie o volumétrica. Por lo tanto, estas representaciones se convierten a menudo en mallas poligonales y en particular en mallas triangulares (véase Figura 1-8(b)). Con este modelo, el objeto es representado por su superficie limitante que está compuesta de un conjunto de caras planas (por lo regular triángulos). De manera más precisa, una malla poligonal contiene un conjunto de puntos 3D (los vértices) que están conectados por aristas para formar un conjunto de facetas poligonales [28]. Existen muchas técnicas de reconstrucción de superficies para producir una malla triangular a partir de una nube de puntos salida de un escáner 3D [29].

Las mallas poligonales pueden representar superficies abiertas o cerradas de una topología arbitraria, con una precisión que depende del número de vértices y facetas. Los algoritmos de intersección, detección de colisiones y de procesamiento son simples y rápidos con este modelo, ya que manipular caras planas es también simple (álgebra lineal). Esta rapidez es particularmente útil en los videojuegos. Estos beneficios hacen que este modelo sea la representación más difundida para los objetos 3D.

El Polígono

Dibujar un polígono es como jugar a conectar los puntos: cada punto es un vértice que define al polígono. Se necesitan al menos tres vértices para definir un polígono. Cada línea del polígono se llama arista (*edge*).

De manera más formal, un polígono puede ser definido como un conjunto de líneas rectas que no se cruzan y que unen puntos coplanarios para encerrar un área única convexa. Un área única significa que el área encerrada no debe dividirse. El requerimiento de que sea convexa quiere decir que dados dos puntos en el polígono, se debe poder dibujar una línea recta que conecte a estos dos puntos sin salirse del área del polígono, como se muestra en la Figura 1-9.

Se requiere que los polígonos sean planos (es decir, que estén definidos en un solo plano). Tres puntos en el espacio definen un plano y por definición, también definen un tipo específico de polígono, el triángulo. Por lo tanto, la mayoría del software de gráficos usa triángulos y tiras

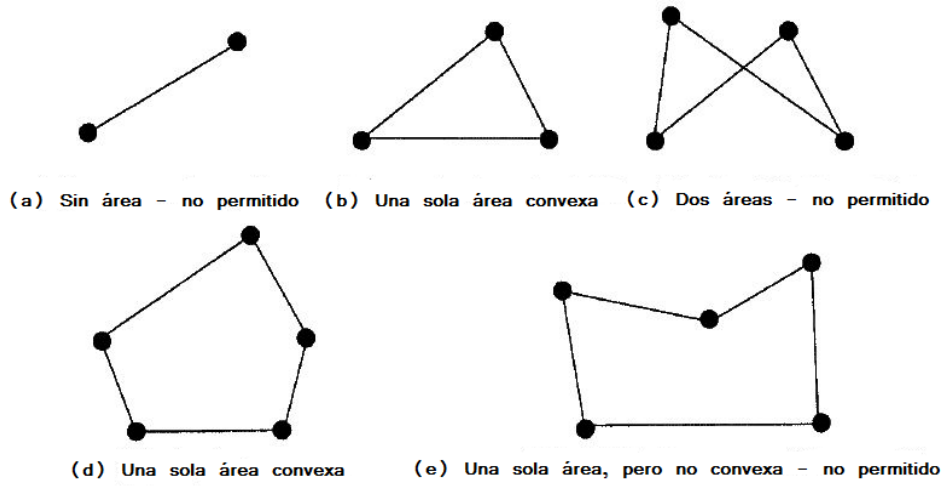

Figura 1-9: Polígonos permitidos y no permitidos.

de triángulos para definir los modelos basados en polígonos.

Polígonos de cara al frente y atrás (Front-Facing and Back-Facing Polygons)

Un polígono tiene dos lados o caras, referidos como la cara al frente y la cara detrás. En CG, el orden en el que se especifiquen los vértices del polígono define su orientación. Por convención, cuando los vértices del polígono son definidos en contra de las manecillas del reloj en la pantalla, se dice que el polígono da la cara al frente. Si se revierte el orden de los vértices, el polígono está mirando hacia atrás, es decir, que la cara de atrás está de frente al observador. La cara del polígono es usada extensamente para definir tratamientos diferentes de las dos caras.

Exclusión de caras ocultas (Back-face Culling)

En la vida real, no se pueden ver todos los lados de un objeto sólido, solamente se ven los lados que están en frente, los lados que miran hacia atrás están obstruidos de nuestra vista. Para lograr este efecto en CG, se hace uso del concepto de polígonos con cara al frente y con cara al fondo. Los polígonos con cara al frente están viendo hacia la cámara y se pintan en la escena. Los polígonos con cara al fondo no están viendo hacia la cámara y por lo tanto son excluidos de la escena. Este proceso se conoce como exclusión de caras ocultas o *back-face culling*. En

la Figura 1-10 se muestra una esfera con todas sus caras visibles y la misma sin las caras que miran al fondo.

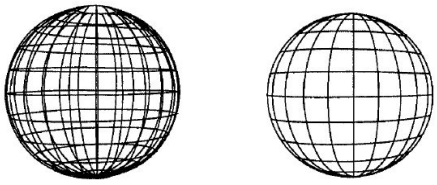

Figura 1-10: Exclusión de caras ocultas aplicada a una esfera.

Mallas poligonales

Una malla poligonal 3D es definida por un conjunto de polígonos planos. Por lo tanto, tal modelo contiene tres tipos de elementos: vértices, aristas y caras. Se pueden agregar atributos adicionales a los vértices, como vectores normales e información de color o textura. Una malla poligonal consiste de dos tipos de información: la geometría y la conectividad. La geometría describe la posición de los vértices en el espacio 3D y la conectividad describe cómo conectar estas posiciones, es decir, la relación entre los elementos de la malla. Estas relaciones especifican, para cada cara, las aristas y los vértices que la componen y para cada vértice, las caras y aristas incidentes. La valencia de un vértice es el número de sus aristas incidentes y el grado de una cara es su número de aristas (véase Figura 1-11).

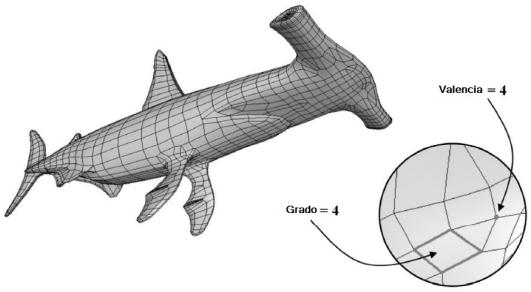

Figura 1-11: Ejemplo de una malla poligonal (2560 vértices, 2562 caras, 5120 aristas), ilustrando la valencia de un vérice y el grado de una cara.

Una malla poligonal es llamada múltiple si cada una de sus aristas pertenece a una o

dos caras (una en el caso de un borde). Las mallas múltiples presentan fuertes propiedades geométricas y por lo tanto son consideradas en la mayoría de los métodos de compresión de malla existentes. Una malla múltiple es asociada con su característica de Euler-Poincaré χ:

$$\chi = v - e + f \tag{1.1}$$

con v, e y f, el número de vértices, aristas y caras de la malla respectivamente. La característica de Euler-Poincaré se relaciona con el género de la superficie correspondiente, de acuerdo a la siguiente ecuación:

$$g = \frac{2c - b - \chi}{2} \tag{1.2}$$

donde c es el número de componentes conectados (aquellos que pertenecen a más de una cara) y b el número de bordes. El género de una superficie describe su complejidad topológica; corresponde al máximo número de curvas cerradas (aquellas que no tienen puntos finales y encierran un área por completo), sin puntos en común, que pueden ser trazadas dentro de esta superficie sin desconectarla. Básicamente el género es el número de asas de la malla; una esfera y un toro son de género 0 y 1 respectivamente [28]. La Figura 1-12 muestra un modelo de género 65.

Figura 1-12: Ejemplo de una malla 3D de género 65 (8830 vértices, 17919 caras).

La representación de modelos por mallas poligonales domina varios formatos de intercambio 3D, tales como VRML[10] y MPEG-4 (2002). El formato de malla estándar es el siguiente: la geometría es representada por una lista de coordenadas indexadas sobre los vértices y la conectividad es descrita por una lista de caras, cada una representada por una lista cíclica

[10] *VRML, Virtual Reality Modeling Language*, es un formato estándar para representar gráficos vectoriales 3D.

de índices de sus vértices incidentes. La Figura 1-13 muestra la representación estándar. Esta representación contiene las coordinadas de los vértices V_0, V_1, V_2, V_3, V_4, V_5 (la geometría) y por cada cara F_0, F_1, F_2, F_3, los índices de los vértices correspondientes (la conectividad).

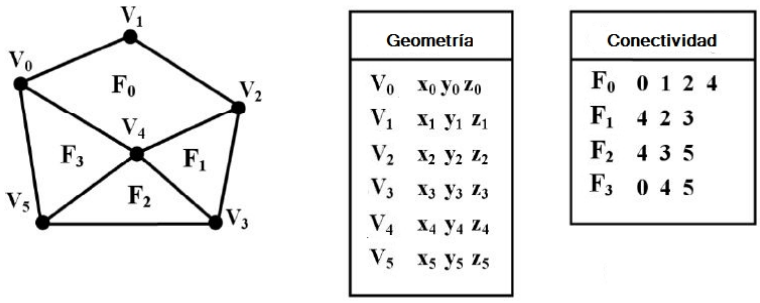

Figura 1-13: Ejemplo de la representación estándar de una malla simple que contiene seis vértices y cuatro caras.

El software de animación 3D llevó a los animadores a una nueva dimensión de posibilidades creativas. A diferencia de la animación en 2D, una vez que un personaje ha sido modelado en 3D, puede ser visto de cualquier ángulo como en la vida real. Grandes esfuerzos se hacen continuamente para mejorar la habilidad de los programas para simular la realidad y lo que puede ofrecerle al animador en cuanto a sus capacidades de modelado, iluminación, uso de texturas, materiales de superficie y procesamiento de las imágenes finales.

En este capítulo se expusieron algunos conceptos relevantes de la animación que sirven de base para el desarrollo del presente trabajo. La técnica de animación por fotogramas clave será utilizada para realizar la animación del rostro. Entender cómo se miden los tiempos en fotogramas y la forma de interpolar entre ellos es importante para entender el proceso de animación del rostro. También, se introdujeron conceptos esenciales del proceso de modelado 3D, como las distintas representaciones de objetos tridimensionales, en especial, la representación de objetos como mallas triangulares. Los modelos 3D del rostro utilizados en este trabajo están representados como mallas triangulares. Por ello, es necesario familiarizarse con la información asociada con este tipo de representación, la información geométrica y de conectividad. El siguiente capítulo se concentra en la animación facial, que aunque parte también de los principios básicos de la animación, merece un estudio por separado debido a su complejidad y singularidad.

Capítulo 2

Metamorfosis de Expresiones Faciales

En el capítulo anterior se presentaron los antecedentes de la animación por computadora y su evolución. En este capítulo se discutirá un área muy particular de la animación: la animación facial.

Es necesario estudiar primero las características anatómicas del rostro para entender la dinámica de una expresión facial. Para ello, se presenta una breve descripción del sistema facial del ser humano y de la generación de las distintas expresiones faciales. Posteriormente, se discuten las diferentes aproximaciones a la animación facial, terminando por describir un poco más a fondo la técnica de metamorfosis, debido a que es la que se usa en el presente trabajo.

2.1. Expresiones Faciales

Nuestra habilidad para identificar expresiones faciales no es algo que tuvimos que aprender, es parte de nuestros instintos. Nuestra maestría para detectar expresiones está tan arraigada que es posible perder la habilidad para reconocer un rostro y poder, aún así, diferenciar una sonrisa de un ceño fruncido. Investigadores que trabajan con gente con lesiones, encontraron un grupo de individuos que podían reconocer varias expresiones incluso cuando no podían reconocer su propio rostro en un espejo. Existe claramente algo fundamental acerca de una habilidad que puede persistir a pesar de tan severo daño cerebral [30].

Tan innata es nuestra capacidad para distinguir expresiones faciales como lo es el reflejo por el cual las expresiones aparecen en primer lugar. No aprendemos a sonreír o llorar al observar cómo lo hacen los adultos. Las expresiones faciales surgen involuntariamente como un estornudo. Un bebé que nace ciego se reirá y llorará como cualquier otro bebé. La mayoría de los expertos cree que las expresiones faciales fundamentales - miedo, alegría, tristesa, sorpresa, disgusto y enojo - son comúnes a todas las sociedades y han permanecido sin cambios por miles de años.

La Estructura de la Cabeza Humana

Las expresiones faciales van y vienen, pasan sobre la superficie de la cara como pequeñas ondas en la superficie de un lago. Las estructuras que se encuentran debajo permanecen sin alterarse, pero las formas óseas de la cabeza hacen que la presencia de las expresiones faciales se sienta en la superficie de manera indirecta. Por ejemplo, los dientes le dan forma a la sonrisa, y el hueso frontal le da forma a un ceño fruncido.

La Estructura Craneofacial

La estructura craneofacial ósea, calavera en términos coloquiales, es el conjunto de huesos del cráneo y los huesos de la cara y es la más importante de las estructuras inferiores que le dan forma al rostro. Las diferencias entre una persona y otra son en gran medida resultado de diferencias en el esqueleto de la cabeza, ya que determina la forma de nuestra cabeza y la localización de nuestras facciones. Kolja Kähler et. al. [31] diseñaron un sistema de reconstrucción facial que se basa en información del esqueleto de la cabeza obtenida de un escáner. Su sistema permite reconstruir un rostro en cerca de una hora a partir de la información de la estructura craneofacial ósea en lugar de semanas, en comparación con el tradicional método de profundidad de tejido manual que se usa en la reconstrucción facial para la identificación post mórtem. Los resultados que obtuvieron demuestran la importancia de la forma de la estructura craneofacial en la forma del rostro, la Figura 2-1 muestra el procedimiento de reconstrucción facial de su sistema.

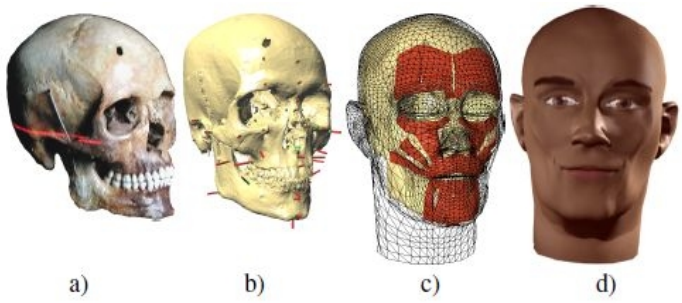

Figura 2-1: Reconstrucción de un rostro a partir de la estructura craneofacial ósea: a) escanear el esqueleto de la cabeza; b) colocar marcas a la malla del esqueleto; c) ajustar la malla de la piel con músculos al esqueleto; d) poner textura a la malla de la piel.

Los Músculos de la Expresión

Nuestros rostros son muy expresivos debido a un grupo complejo de pequeños músculos. Hay una red de estos músculos debajo de la superficie de la cara y con sus movimientos pueden alterar por completo la apariencia del rostro. Aunque estos músculos son relativamente débiles y pequeños, están tan adheridos a la superficie de la piel que un movimiento de fibra muscular modesto frecuentemente se traduce en un gran movimiento en la piel. Algunos de estos movimientos musculares se reconocen como reflejos de estados emocionales, dando como resultado las expresiones faciales.

La forma en que los músculos faciales están mezclados con todo lo demás debajo de la piel les hizo la vida difícil a los primeros anatomistas que trataban de trazar los músculos de la cara. La famosa anatomía de Vesalius, publicada a finales del siglo XVI, muestra los músculos faciales en una forma vaga y engañosa. Otros sistemas musculares del cuerpo, más grandes y fáciles de analizar, fueron retratados con más exactitud.

Casi un siglo antes, otro anatomista pionero había explorado pacientemente los músculos faciales e hizo dibujos exactos de ellos. Pero los dibujos anatómicos de Leonardo da Vinci, como la mayoría de su trabajo científico y artístico, para el tiempo de Vesalius estaba disperso en colecciones privadas y desconocido para el mundo en general. El resultado del meticuloso esfuerzo de Leonardo no fue solamente sus dibujos anatómicos. Los hombres en sus escenas de batalla, así como las mujeres en sus retratos, tienen rostros más realistas y con más vida, que

cualquier otro que haya aparecido anteriormente en la pintura.

Aunque Leonardo y otros habían trazado los músculos faciales, la función de los distintos músculos no era entendida del todo hasta el siglo diecinueve. A mediados del siglo diecinueve, Duchenne de Boulogne encontró que leves sacudidas eléctricas en varios puntos de la cara causaban que los músculos se contrajeran individualmente. Sus fotografías de sonrisas y gruñidos inducidos por electricidad son a la vez extrañas e impulsivas; su descripción de qué músculo hace qué, fue un avance importante.

La pregunta de por qué reímos y gruñimos fue emprendida por un hombre más famoso por su trabajo en otro campo. El libro de Charles Darwin de expresiones faciales, *The Expression of Emotion in Man and Animals* (1872), sigue siendo probablemente el mejor libro en la materia.

Los músculos faciales son diferentes de los otros músculos del cuerpo en lo que hacen y cómo lo hacen. La mayoría de nuestros músculos, como los bíceps en el brazo o los tendones de la corva en la pierna, se estiran de un hueso a otro, usualmente a través de una articulación. Cuando estos músculos se contraen, los huesos involucrados se acercan, a menudo doblándose en una articulación. Los músculos siempre jalan cosas para acercarlas, es la única manera en la que funcionan. Se encojen y sus extremos se aproximan, nunca empujan. Por ejemplo, cuando los bíceps se contraen, el radio en el antebrazo es acercado al húmero en el brazo superior, y el brazo se dobla. Cuando los tendones de la corva se contraen, la tibia se aproxima al fémur, y la pierna se dobla.

Los músculos faciales, sin embargo, usualmente tienen sólo un extremo fijo, atado directamente o indirectamente a un hueso de la estructura craneofacial ósea. El otro extremo del músculo está cosido en la piel (o en otro músculo atado a la piel). Cuando un músculo de expresión se contrae, la piel, en lugar del hueso, se mueve. La porción de piel cerca de la tira del extremo de los músculos es jalada en la dirección de su atadura al hueso.

Por ejemplo el músculo de la sonrisa, el cigomático mayor. Tiene un extremo atado al pómulo, justo abajo de la esquina exterior del ojo. Entonces el músculo se estira diagonalmente hacia abajo en dirección de la boca, donde está atado indirectamente a su esquina exterior. Cuando se contrae, la esquina de la boca sube en dirección del pómulo, y sonreímos. Así funcionan la mayoría de los músculos de expresión [30].

La Figura 2-2 muestra los músculos encargados de la expresión facial. De los más de 26

músculos que mueven la cara, solamente los que se muestran son responsables de la expresión facial.

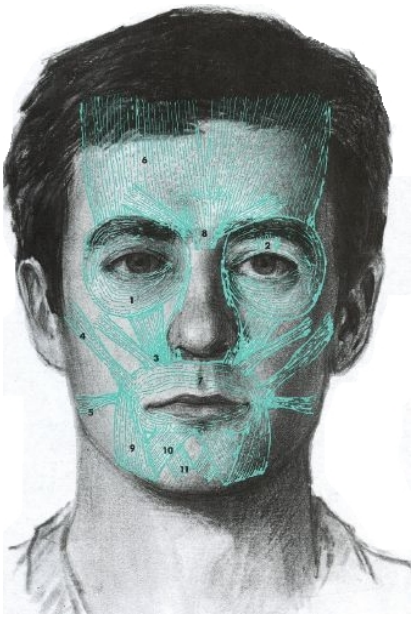

Figura 2-2: Músculos clave para la expresión facial: 1. Orbicularis oculi, 2. Levator palpabrae, 3. Levator labii superioris, 4. Zygomatic Major, 5. Risorius/platysma, 6. Frontalis, 7. Orbicularis oris, 8. Corrugator, 9. Triangularis, 10. Depressor labii inferioris, 11. Mentalis.

El orbicularis oculi u orbicular de los párpados, está atado a la órbita interior y la piel de la mejilla, aprieta el ojo, como al bizquear. El levator palpabrae se origina en la órbita y se ata al párpado superior, sube el párpado, como al sorprenderse. El levator labii superioris o elevador propio del labio superior, tiene tres ramas, la rama interna se origina en la base de la nariz; la rama intermedia en el borde inferior de la órbita; la rama exterior en el arco cigomático. Todas se insertan en la piel arriba del labio superior. El zygomatic major o cigomático mayor, que se origina en el arco cigomático y se inserta en la esquina de la boca, jala la boca para sonreír. El risorius o músculo risorio se origina en la parte posterior de la quijada y se inserta en la esquina de la boca; el platysma se origina en el pecho superior y se inserta también en la esquina de la boca. Ambos trabajan juntos, estirando la boca, como al llorar (el platysma no se muestra en la figura, ya que cubre otros músculos). El músculo frontalis se origina cerca de la tapa del cráneo y se inserta debajo de las cejas, sube las cejas como al sorprenderse. El orbicularis oris

se origina de los músculos en la esquina de la boca, encrezpa y aprieta los labios. El músculo corrugador se origina en el puente nasal y está atado a la piel debajo de la mitad de la ceja, baja el extremo interior de la ceja. El triangularis se origina sobre el margen inferior de la quijada y se inserta en la esquina de la boca, tirándola hacia abajo. El depressor labii inferioris se origina sobre la parte inferior de la barbilla y se inserta en el labio inferior, lo tira hacia abajo como al hablar. El músculo mentalis se origina justo abajo de los dientes, en la quijada inferior y se inserta en la piel de la barbilla, empuja el labio inferior hacia arriba.

2.2. Animación Facial

La animación y modelado facial se refieren a técnicas para representar el rostro humano gráficamente en un sistema computacional y animar tal rostro de manera consistente con los humanos reales. Ésta es a menudo considerada una de las tareas más retadoras en el campo de la animación, ya que todos los humanos somos muy hábiles en identificar movimientos faciales innaturales, la más ligera inconsistencia alerta al espectador y la animación pierde su realismo [32]. Durante sus casi cuarenta años de existencia, la animación facial ha visto la invención de una multitud de tecnologías y luego desaparecer por obsoletas. Los métodos de modelado y animación han sido en su mayoría dictados por el hardware disponible, el cual evolucionó de gran manera a través de los años. Muchas de las ambiciones en la animación que han sido consideradas sueños en su momento ahora son una realidad a medida que se descubren y exploran nuevas fronteras.

2.2.1. Tipos de Modelos Faciales

El primer problema al que se enfrenta un animador en la animación facial es el de crear la geometría del modelo facial para hacerla apropiada para la animación. Los modelos en la animación facial varían mucho, desde una geometría simple hasta los basados en la anatomía. Generalmente, la complejidad es dictada por el uso que se propone. Cuando se está decidiendo sobre la construcción del modelo, algunos factores importantes a considerar son el método de adquisición de la información, el control de movimiento y su correspondiente método de adquisición, la calidad del procesamiento de las imágenes finales y la calidad de movimiento.

El primer factor tiene que ver con el método por el cual se obtiene la geometría de la cabeza del sujeto o personaje. El segundo factor tiene que ver con el método por el cual se obtiene la información que describe los cambios en la geometría. La calidad de la imagen final con respecto a la suavidad de la superficie y sus atributos, es la tercera preocupación. Por último, se debe tomar en cuenta la calidad del movimiento computado.

Se puede hablar del modelo en términos de sus propiedades estáticas o sus propiedades dinámicas. Las estáticas tratan con la geometría del modelo en su forma neutral, mientras que las dinámicas tienen que ver con la deformación de la geometría del modelo durante la animación. Se han usado tres métodos principales para tratar la geometría del modelo. Los modelos poligonales son frecuentemente los más usados por su simplicidad; los splines se usan cuando se desea una superficie suave. Actualmente, también se usan superficies de subdivisión con un éxito moderado.

Los modelos poligonales son relativamente fáciles de crear y deformar. Sin embargo, la suavidad de la superficie está relacionada directamente con la complejidad del modelo y los modelos poligonales son visualmente inferiores a otros métodos de modelado de la superficie facial. Actualmente, los métodos de adquisición de información muestrean solamente la superficie, produciendo información discreta y posteriormente, se aplican técnicas de ajuste de superficies.

Los modelos por splines usan típicamente parches bicúbicos de cuatro lados, tales como el Bezier o B-spline, para representar el rostro, mientras que los parches de superficie ofrecen la ventaja de baja complejidad de información en comparación con las técnicas poligonales para generar superficies suaves, tienen varias desventajas cuando se trata de modelar un objeto como el rostro. Con la tecnología estándar de parches de superficie, se utiliza una malla rectangular de puntos de control para modelar el objeto completo. Como resultado, es difícil mantener una baja complejidad de información al incorporar pequeños detalles y facciones angulosas, ya que filas completas o columnas completas de información de control deben ser modificadas. Por lo tanto, una pequeña añadidura a un área local de la superficie para representar mejor una facción significa que se debe añadir información a través de toda la superficie.

Los B-splines jerárquicos, introducidos por Forsey y Bartels [33], son un mecanismo por el cual, detalles locales pueden ser agregados a una superficie B-spline evitando modificaciones globales requeridos por los B-splines estándares. Se colocan finos puntos de control para res-

olución sobre la superficie mientras que la continuidad se mantiene cuidadosamente. De esta forma, se pueden agregar detalles a la superficie. La organización es jerárquica, así que se pueden añadir detalles cada vez más finos.

Las superficies de subdivisión (por ejemplo, [34]) usan una malla poligonal de control que está refinada, en el límite, a una superficie suave. El refinamiento puede ser finalizado a una resolución intermedia y procesado como una malla poligonal. Las superficies de subdivisión tienen la ventaja de ser capaces de crear una complejidad local sin una global. Proveen una interfaz interactiva de fácil uso para desarrollar nuevos modelos. Sin embargo, son difíciles para interpolarse a un conjunto específico de información, lo que hace que modelar un rostro en específico sea problemático.

Las superficies definidas implícitamente también han sido usadas para modelar rostros, pero tales modelos, por lo regular, se vuelven muy complejos cuando el animador trata con detalles pequeños.

2.2.2. Crear el Modelo

Crear el modelo de una cabeza humana desde cero no es fácil. No solamente se debe generar la forma correcta, sino que cuando el objetivo es la animación facial, los elementos de la geometría (vértices, aristas) deben ser colocados apropiadamente para poder controlar con precisión el movimiento de la superficie. Si el modelo es denso en el número de elementos usados, entonces la colocación se vuelve menos preocupante, pero en modelos de una resolución relativamente baja puede ser un problema. Por supuesto, una aproximación es usar un sistema CAD[1] que permita al usuario construir el modelo. Mientras que esta aproximación le da al artista la mayor libertad requiere mucha habilidad.

Además de este método, existen dos métodos principales para crear modelos faciales: digitalización usando alguna referencia física y modificación de un modelo existente. El primero es útil cuando se desea el modelo de una persona en particular; el segundo cuando el control de la animación ya está incluido en el modelo genérico.

Los modelos también se pueden generar a partir de fotografías. Esto tiene la ventaja de que no se requiere la presencia de un modelo físico una vez que la fotografía se ha tomado y tiene

[1] *Computer Aided Design.*

aplicaciones en la videoconferencia y la compresión. Aunque la mayoría de las aproximaciones fotográficas modifican un modelo existente al localizar rasgos, un método común para generar un modelo desde cero es tomar fotos de frente y de lado de una cara sobre la cual se ha dibujado una malla (Figura 2-3). Las correspondencias entre puntos se pueden establecer entre las dos imágenes ya sea de manera interactiva o localizando rasgos automáticamente y la malla se puede reconstruir en el espacio tridimensional. Ya que se asume un rostro simétrico, solamente se requiere uno de los lados y la mitad del rostro de frente.

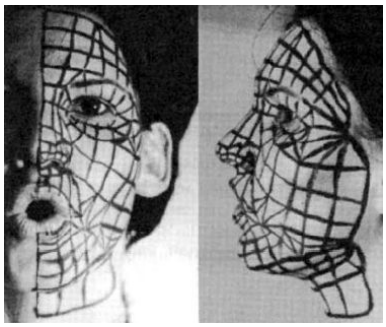

Figura 2-3: Fotografías a partir de las cuales una cara puede ser digitalizada.

2.2.3. Técnicas de Animación Facial

Las aproximaciones a la animación facial pueden ser agrupadas en dos: aquellas basadas en la manipulación de la geometría y las basadas en manipulación de imágenes. En esta sección se discutirán las técnicas para manipular la geometría ya que son las que nos interesan, las técnicas de manipulación de imágenes se pueden apreciar en el trabajo de Tony Ezzat y Tomaso Poggio [35].

Animación Facial por Fotogramas Clave

La aproximación más simple a la animación facial es definir un conjunto de poses clave. La animación facial se produce al seleccionar dos de estas poses clave e interpolar entre las posiciones de los vértices correspondientes en ellas. Esto obliga a que los movimientos disponibles sean la interpolación de una pose clave a otra. Para generalizar un poco más, se puede utilizar una suma ponderada de dos o más poses clave en la cual los pesos suman uno. Cada posición de

un vértice es calculada como una combinación lineal de su posición correspondiente en cada una de las poses que tienen un peso distinto de cero. Esto puede utilizarse para producir poses faciales que no están directamente representadas por las claves. Sin embargo, esto es todavía bastante restrictivo porque las diversas partes del modelo facial no son contraladas individualmente por el animador. La animación está todavía restringida a las poses representadas como una combinación lineal de las claves.

Modelos con Parámetros

El Sistema de Codificación de Acciones Faciales o FACS (*Facial Action Coding System*) es el resultado de la investigación que condujeron los psicólogos Ekman y Friesen [36], con el objetivo de descomponer todas las expresiones faciales en un conjunto de movimientos faciales básicos. Estos movimientos, llamados Unidades de Acción o AUs (*Action Units*), podían usarse para describir todas las expresiones faciales al considerar sus combinaciones.

El dar parámetros a un modelo facial de acuerdo a acciones primitivas y luego controlar los valores de estos parámetros sobre el transcurso del tiempo, es una de las formas más comunes para implementar la animación facial. De manera abstracta, cualquier contorsión facial imaginable puede ser considerada como un punto en un espacio de dimensión n de todas las posibles posturas faciales. Una parametrización de un espacio debe tener cobertura total y debe ser fácil de usar. Cobertura total significa que el espacio alcanzable por las combinaciones (lineales) de los parámetros incluye todos (o por lo menos la mayoría) de los puntos interesantes en el espacio. Por supuesto, la definición de la palabra *interesante* puede variar entre aplicaciones, así que una parametrización general útil incluye tanto del espacio como sea posible. Para que una parametrización sea fácil de usar, el conjunto de parámetros debe ser lo más pequeño posible, el efecto de cada parámetro debe ser independiente del efecto de cualquier otro y el efecto de los parámetros debe ser intuitivo. Por supuesto, en algo tan complejo como la animación facial, lograr todos estos objetivos es probablemente imposible, así que determinar compensaciones apropiadas es una actividad importante al diseñar la parametrización. El modelo parametrizado más popular se acredita a Parke [37].

Modelos Musculares

Los modelos con parámetros codifican el desplazamiento de la piel en términos de un valor de un parámetro arbitrario. Los modelos basados en músculos son más sofisticados. Existen por lo regular tres tipos de músculos que necesitan modelarse para el rostro: lineales, láminas y esfínter. El modelo lineal es un músculo que se contrae y jala un punto (el punto de inserción) hacia otro (el punto de unión). El modelo de lámina actúa como un arreglo paralelo de músculos y tiene una línea de unión en cada uno de sus dos extremos en lugar de un solo punto de unión como el modelo lineal. El esfínter se contrae radialmente hacia un centro imaginario. El usuario especifica directa o indirectamente la actividad muscular a la cual el modelo facial reacciona. Hay tres aspectos que diferencian un modelo muscular de otro: la geometría del arreglo músculo-piel, el modelo de piel usado y el modelo de músculos usado.

La principal característica distintiva del arreglo geométrico de los músculos es si están modelados en la superficie de la piel o si están unidos a una capa estructural debajo de la piel (por ejemplo, hueso). El primer caso es más simple ya que solamente se necesita el modelo de la superficie del rostro para el sistema de animación (Figura 2-4(a)). El segundo caso es más correcto en cuanto a la anatomía y por lo tanto promete resultados más exactos, pero requiere una mayor estructura geométrica en el modelo y es, por consiguiente, más difícil de construir (Figura 2-4(b)).

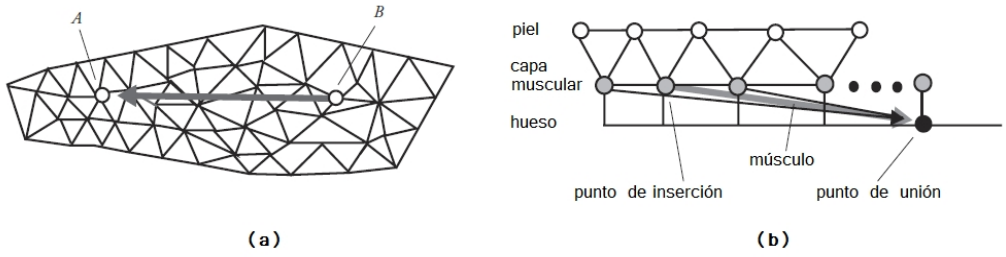

Figura 2-4: Modelos musculares: (a) Parte de la superficie geométrica del rostro que muestra el punto de unión (A) y el punto de inserción (B) de un músculo lineal; el punto B es jalado hacia el punto A. (b) Sección transversal de un músculo de tres capas como el presentado por Parke y Waters, el músculo sólo afecta directamente los nodos en la capa de músculos.

El modelo usado para la piel dictará la forma en que el área alrededor del punto de inserción

de un músculo (lineal) reacciona cuando se activa el músculo; el punto de inserción se moverá cierta distancia determinada por el músculo. La forma en que la deformación se propaga sobre la piel, como resultado de la activación de este músculo, determina qué tan elástica o plástica parecerá la superficie. El modelo más simple está basado en la distancia geométrica desde el punto y la desviación a partir del vector muscular. Por ejemplo, el efecto del músculo puede atenuarse basado en la distancia a la que se encuentra un punto dado del punto de inserción y en el ángulo de desviación del vector de desplazamiento del punto de inserción. Un modelo de la piel un poco más sofisticado representa cada arista de la geometría de la piel como un resorte y controla la propagación de la deformación basado en constantes del resorte. El punto de inserción es movido por la acción del músculo y este desplazamiento crea fuerzas de restauración en los resortes unidos al punto de inserción, que mueven los vértices adyacentes, los que a su vez mueven los vértices que están unidos a ellos y así sucesivamente (Figura 2-5(a)).

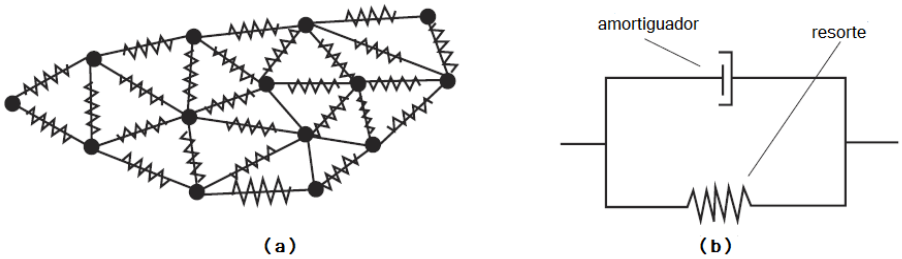

Figura 2-5: Modelos de piel. (a) Malla de resortes como el modelo de piel; el desplazamiento del punto de inserción se propaga a través de la malla de acuerdo a las fuerzas impartidas por los resortes. (b) Modelo visco-elástico de Voight; el movimiento inducido por las fuerzas de los resortes es amortiguado.

El modelo de Voight es más complicado, trata a la piel como un elemento visco-elástico al combinar un resorte y un amortiguador en paralelo (Figura 2-5(b)). El movimiento inducido por el resorte es amortiguado como una función del cambio de longitud de la arista.

El modelo de músculo determina la función que se utiliza para computar la contracción del músculo. Las alternativas para el modelo de músculo son similares a las de la piel, con la diferencia de que los músculos son elementos activos, mientras que la piel está compuesta de elementos pasivos. Al usar un músculo lineal por ejemplo, el desplazamiento del punto de inserción es producido como resultado de la activación del músculo. Los modelos simples para el

músculo simplemente especifican el desplazamiento del punto de inserción basados en la cantidad de activación. Los modelos de músculo más exactos físicamente calculan el efecto de las fuerzas musculares. El modelo dinámico más simple usa un resorte para representar al músculo. Activar el músculo resulta en un cambio de su longitud de reposo para inducir una fuerza en el punto de inserción. Los modelos musculares más sofisticados incluyen efectos de amortiguamiento.

Sifakis et al. [38] construyeron un modelo muscular facial muy exacto, usando los principios derivados de los principios más generales de construcción de músculos de Teran et al. [39]. Su técnica está basada en algoritmos de elemento finito. Una característica novedosa del modelo es que su acción muscular puede interaccionar con el ambiente, es decir, las fuerzas musculares pueden combinarse con fuerzas externas tales como colisiones, produciendo como resultado el efecto que se muestra en la Figura 2-6.

Figura 2-6: Interacción de la cara con un objeto externo.

2.3. Metamorfosis (Morphing)

La metamorfosis, también conocida como morphing, es la transformación gradual de una figura (el origen) a otra (el destino). La metamorfosis tiene un amplio uso práctico en áreas como los gráficos por computadora, la animación y el modelado. Actualmente, para lograr resultados más espectaculares y exactos, el proceso de metamorfosis requiere que mucho del trabajo se haga manualmente. Un reto importante de investigación es desarrollar técnicas que automaticen este proceso lo más que se pueda.

El problema de metamorfosis ha sido investigado en muchos contextos. Por ejemplo, con imágenes de dos dimensiones [41], polígonos [42] e incluso representaciones volumétricas basadas en

voxels [43]. El proceso de metamorfosis siempre consiste en resolver dos problemas principales. El primero es encontrar una correspondencia entre los elementos de las dos representaciones de las figuras. El segundo problema es encontrar trayectorias que puedan recorrer los elementos correspondientes durante el proceso de metamorfosis. Desgraciadamente, no existe una definición formal de una correspondencia exitosa ni una definición de una solución exitosa para el problema de las trayectorias.

Una solución ingenua al problema de las trayectorias, una vez que se ha establecido una correspondencia, es hacer que las trayectorias sean líneas rectas. Una metamorfosis generada de esta forma se dice que es una metamorfosis lineal. Desafortunadamente, esta simple aproximación puede llevar a resultados indeseados. Las figuras intermedias pueden desaparecer, es decir, degenerarse a un solo punto. Más aún, las figuras intermedias pueden intersecarse entre ellas mismas, incluso si las figuras origen y destino no lo hacen [40].

En cuanto a la correspondencia entre los elementos de la figura origen y la figura destino, el proceso depende de la representación de éstas. Según la representación de los objetos a ser transformados se puede hablar de metamorfosis 2D y metamorfosis 3D. La metamorfosis 2D se aplica a imágenes digitales, que pueden ser fotografías del mundo real, o bien imágenes generadas sintéticamente. La metamorfosis 3D trabaja con objetos con representación tridimensional, los cuales necesariamente se proyectan a 2D para que puedan ser presentados en la pantalla.

2.3.1. Interpolación

El principio de la mayoría de la animación es la interpolación de valores. Uno de los ejemplos más sencillos de animación es la interpolación de la posición de un punto en el espacio. Cuando una curva es construida a partir de un conjunto de puntos y la curva pasa por todos ellos, se dice que interpola los puntos. El término *interpolación* también se usa por lo general para referirse a todas las aproximaciones para construir una curva a partir de un conjunto de puntos. Existen muchas técnicas de interpolación como la interpolación Hermite, o los splines Catmull-Rom entre otras, aquí se discute solamente la interpolación lineal simple que es la de interés para este trabajo.

Interpolación Lineal Simple

La interpolación lineal simple está dada por la ecuación 2.1 y se muestra en la Figura 2-7. Note que los factores, $1 - u$ y u suman uno, y además, $0 \leq u \leq 1$. Esta propiedad garantiza que la curva interpolada (en este caso una línea recta) cae dentro del casco convexo (*convex hull*) de las entidades geométricas que están siendo interpoladas (en este caso sencillo el casco convexo es la misma línea recta).

$$P(u) = (1 - u) \cdot P0 + u \cdot P1 \tag{2.1}$$

La ecuación para la interpolación lineal también puede escribirse como en la ecuación 2.2. Esta forma es típica en ecuaciones polinomiales en donde los términos son agrupados de acuerdo a los coeficientes de la variable elevada a alguna potencia.

$$P(u) = P0 + u \cdot (P1 - P0) \tag{2.2}$$

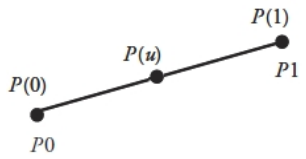

Figura 2-7: Interpolación lineal.

2.3.2. Metamorfosis 2D

La metamorfosis de imágenes de dos dimensiones ha llegado a conocerse como morphing. Por lo regular, el usuario está interesado en transformar una imagen, llamada la imagen origen, en otra imagen, llamada la imagen destino. Hay varias técnicas en la literatura para especificar y efectuar la transformación. La tarea principal es que el usuario especifique los elementos correspondientes en las dos imágenes; estas correspondencias son usadas para controlar la transformación. Aquí se presentan dos aproximaciones. La primer técnica está basada en una malla de coordenadas definidas por el usuario sobrepuesta en cada imagen. Esta malla impone un

espacio de coordenadas para relacionar una imagen a la otra. La segunda técnica está basada en líneas de características también definidas por el usuario, una en cada imagen. Las líneas marcan características correspondientes en las dos imágenes.

La Aproximación por Malla de Coordenadas

Para transformar una imagen en otra, el usuario define una malla curvilínea sobre cada una de las dos imágenes sobre las que se aplicará la metamorfosis. Es responsabilidad del usuario definir las mallas de tal forma que los elementos correspondientes en las imágenes se encuentren en las celdas correspondientes de las mallas. El usuario define la malla localizando el mismo número de puntos de intersección de la malla en ambas imágenes; la malla debe ser definida en el borde de las imágenes para incluirla en su totalidad. Una malla curva se genera usando los puntos de intersección como puntos de control para un esquema de interpolación. La Figura 2-8 muestra un ejemplo de definición de mallas para la metamorfosis [44].

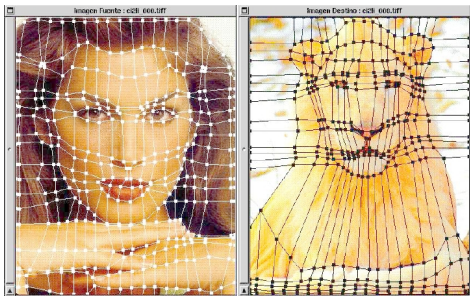

Figura 2-8: Ejemplo de la definición de mallas de coordenadas para transformar una mujer en un león.

Para generar una imagen intermedia entre la imagen origen y la imagen destino, digamos t $(0 < t < 1{,}0)$, los vértices (los puntos de intersección de las curvas) de las mallas origen y destino son interpolados para formar una malla intermedia. Esta interpolación se puede hacer linealmente, o las mallas de fotogramas clave adyacentes se pueden usar para llevar a cabo una interpolación de orden más alto. Los pixeles de la imagen origen y destino son extendidos y comprimidos de acuerdo a la malla intermedia generando versiones deformadas tanto de la imagen origen como de la imagen destino. Después se disuelve de manera cruzada píxel por píxel entre las dos imágenes deformadas para generar la imagen final.

Metamorfosis Basada en Características

En lugar de usar una malla de coordenadas, el usuario puede establecer la correspondencia entre las imágenes usando líneas de características (*feature lines*). Las líneas son dibujadas sobre las dos imágenes para identificar catacterísticas o rasgos que corresponden entre sí; las líneas de características son interpoladas para formar un conjunto intermedio de líneas de características. La interpolación puede ser basada ya sea en interpolar puntos de los extremos o interpolar puntos centrales y su orientación. En cualquier caso, se establece una proyección de cada píxel en la imagen intermedia a cada línea de características interpolada y se calcula un peso relativo que indica qué tanto influye esa línea de características sobre el píxel. La proyección es usada en la imagen origen para localizar el píxel en esta imagen que corresponde al píxel en la imagen intermedia. El peso relativo es usado para promediar las ubicaciones en la imagen origen generadas por líneas de características múltiples obteniendo una ubicación final en la imagen origen. Esta ubicación es usada para determinar el color del píxel en la imagen intermedia. Este mismo procedimiento es usado en la imagen destino para formar su imagen intermedia. Estas imágenes intermedias se disuelven de manera cruzada para formar la imagen intermedia final (Figura 2-9).

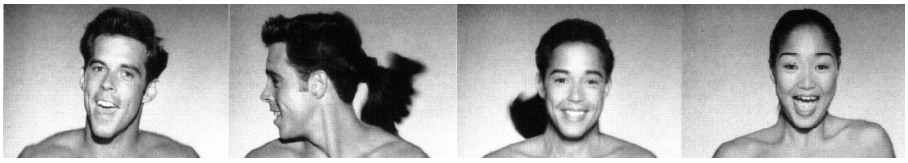

Figura 2-9: Secuencia del video de Michael Jackson, *Black or White*, donde se usó la metamorfosis por líneas de características.

2.3.3. Metamorfosis 3D

Transformar un objeto 3D en otro objeto 3D es un efecto útil, pero uno que tiene problemas para los cuales las soluciones de propósito general siguen en desarrollo. Existen varias soluciones con sus ventajas y desventajas. Las técnicas caen dentro de dos categorías: basadas en superficie y basadas en volumen. Las técnicas basadas en superficie usan la representación de límite de los objetos y modifican los vértices, las aristas o ambos, para que las topologías vértice-arista de los

dos objetos se igualen. Una vez que esto se logra, los vértices del objeto pueden ser interpolados uno a uno. Las técnicas basadas en superficie usualmente tienen algún tipo de restricción acerca del tipo de objetos que pueden manejar, especialmente los objetos que tienen agujeros. El número de agujeros a través de un objeto es un atributo importante de la estructura del objeto, o topología. Las técnicas basadas en volumen consideran el volumen contenido en los objetos y mezclan un volumen en otro. Estas técnicas tienen la ventaja de ser menos sensibles a topologías diferentes entre los objetos. Sin embargo, las técnicas basadas en volumen usualmente requieren representaciones volumétricas de los objetos y por lo tanto tienden a ser computacionalmente más intensivas que las aproximaciones basadas en superficie.

Técnicas para representaciones basadas en superficie

Las representaciones basadas en superficie son muy populares para representar objetos 3D y mundos virtuales 3D. Una gran cantidad de modelos y estructuras de datos han sido propuestos, pero las superficies poligonales y las superficies parametrizadas son las más usadas. Las estructuras de datos casi siempre contienen información de la topología y de la geometría. Sin embargo, este tipo de representación es muy restringida y rígida, lo que significa que si se cambia al azar aunque sea un par de valores en la estructura de datos, puede resultar un objeto no representable. Es por eso que llevar a cabo la metamorfosis se vuelve más complejo. La presencia de topología y geometría en una representación basada en superficie divide a la metamorfosis en dos pasos:

- Establecer una correspondencia entre el objeto fuente y el destino.
- Interpolar las posiciones (la geometría) de las características correspondientes.

Topologías Iguales. El caso más simple para transformar un objeto en otro es cuando los dos objetos a ser interpolados comparten la misma topología vértice-arista. Aquí, los objetos son transformados meramente al interpolar las posiciones de los vértices uno por uno. La correspondencia entre los dos objetos se establece por la estructura de conectividad vértice-arista que comparten los dos objetos. El problema de la interpolación se soluciona al interpolar las posiciones 3D de los vértices.

Transformación de Forma para Poliedros. El algoritmo de James R. Kent et al. [45] transforma entre sí las estructuras topológicas de dos modelos 3D que sean poliedros en forma de estrella[2] (star-shaped) y que además sean de género 0, es decir, que no tengan agujeros. La principal contribución de este método es la correspondencia entre vértices. El algoritmo une las estructuras topológicas en una malla común de vértices, bordes y caras. Se generan dos nuevos modelos equivalentes topológicamente y con la misma forma de los originales, pero que permiten transformaciones de uno a otro que se pueden procesar fácilmente.

La superficie de los dos objetos sólidos se proyecta sobre una esfera. Esta proyección se usa para identificar las correspondencias entre los puntos en los dos objetos originales al asociar pares de puntos que se proyecten a la misma ubicación en la esfera, esto establece una correspondencia uno a uno entre los puntos en la superficie de los dos objetos.

El algoritmo de correspondencia para poliedros sólidos de género 0 sigue los siguientes pasos:

- Se reciben como entrada dos modelos 3D que cumplen las siguientes características: son de género 0 y son poliedros en forma de estrella.
- Se proyecta la topología de los dos modelos sobre una esfera.
- Se entremezclan las dos topologías al unir las caras proyectadas de un modelo a las caras proyectadas del otro.
- La topología combinada se proyecta entonces sobre la superficie de los dos modelos originales.
- Como salida se obtienen dos modelos que tienen la misma forma que los dos modelos originales, pero que comparten una topología común. Esto permite que una transformación entre las dos formas se procese fácilmente al interpolar las coordenadas de cada par de vértices correspondientes.

Hay varios métodos para proyectar los modelos sobre la esfera, sólo deben cumplir dos condiciones: que sean uno a uno y que sean continuos. La metamorfosis ocurre al interpolar

[2] Un polígono en forma de estrella (2D) es aquel en el que hay por lo menos un punto a partir del cual se puede dibujar una línea a cualquier punto en el contorno del polígono sin intersecar el contorno; un poliedro en forma de estrella (3D) se define de manera similar. El conjunto de puntos a partir de los cuales se puede ver el contorno entero es referido como el núcleo (kernel) del polígono (en el caso 2D) o el poliedro (en el caso 3D).

las posiciones de los vértices correspondientes y cada vértice se mueve radialmente sobre un rayo que sale desde un punto central. El usuario debe indicar ese punto central y seleccionar la orientación relativa de los dos objetos. Además de la interpolación lineal, se usan splines Hermite para la trayectoria de cada vértice, con los vectores tangentes de la spline igual a las normales de los vértices. En la Figura 2-10 se muestra un ejemplo del algoritmo de correspondencia.

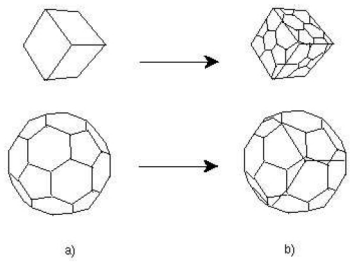

Figura 2-10: Un ejemplo del algoritmo de correspondencia. (a) Los modelos originales, (b) modelos con la topología combinada proyectada en la superficie.

Metamorfosis de Poliedros basada en Características. Francis Lazarus y Anne Verroust [46] presentan una técnica para transformar dos poliedros. Ellos tratan el problema de la transformación como un todo, no disocian la correspondencia de características y la interpolación, aunque sí resuelven cada uno de los dos problemas. El usuario debe especificar una curva 3D para cada objeto, la cual va a hacer el papel de un eje, de tal forma que los objetos tengan forma de estrella alrededor de sus ejes respectivos. Otra opción es que el usuario especifique un punto para procesar un eje candidato. Este eje se usa para construir una malla cilíndrica que se aproxime al objeto. Entonces la metamorfosis se lleva a cabo con base en la interpolación de las mallas de los dos objetos, compuesta con una interpolación radial de cada punto de la malla.

En la Figura 2-11 se muestran los resultados obtenidos por Lazarus y Verroust al aplicar la metamorfosis de poliedros basada en características. El objeto fuente es un cisne y el objeto destino es un piano.

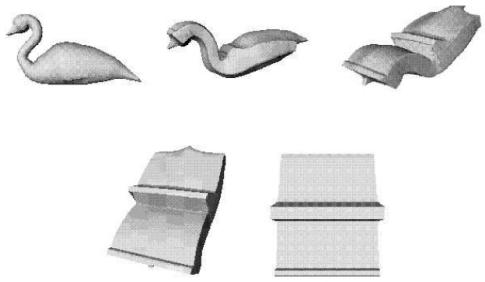

Figura 2-11: Metamorfosis de poliedros basada en características.

Metamorfosis con Descomposición basada en Características.

La técnica descrita por Arthur D. Gregory [47] se aplica a dos poliedros con topologías equivalentes. La especificación de correspondencias consiste en que para cada arco de la malla el usuario debe seleccionar un par de puntos en cada superficie. Entonces se procesa una línea geodésica entre cada par para formar un arco de la malla. En base a estas mallas de características, el algoritmo descompone la superficie de los poliedros en parches, llamados parches de metamorfosis (la misma cantidad para ambos poliedros), se realiza una proyección para cada parche de metamorfosis a un polígono, se fusionan y se construye un poliedro fusionado cuya conectividad topológica es una combinación de la conectividad topológica de los poliedros de entrada. El resultado de haber procesado la correspondencia entre los poliedros, es un poliedro fusionado con la topología de ambos modelos de entrada, para el que cada vértice tiene una ubicación en los dos modelos de entrada.

Figura 2-12: Metamorfosis con descomposición basada en características de una taza en una dona.

Ahora cada vértice del poliedro fuente tiene un vértice correspondiente en el poliedro destino. Durante la metamorfosis, los vértices viajan de sus posiciones en el objeto fuente a sus posiciones respectivas en el objeto destino a lo largo de trayectorias de metamorfosis. Estas

trayectorias se representan como curvas de Bézier. El sistema interpola las trayectorias para el resto de los vértices del poliedro fusionado usando contribuciones de peso de las trayectorias de metamorfosis en los vértices extremos. Además se interpolan otros atributos, como colores de vértices, coeficientes de iluminación, vectores normales, etc. La interpolación de estos atributos de superficie ocurre durante los pasos de proyección y fusión. La Figura 2-12 muestra el resultado de transformar una dona en una taza, mediante la metamorfosis con descomposición basada en características.

Técnicas para Representaciones basadas en Volumen

Estas técnicas son las preferidas para las aplicaciones de metamorfosis debido a que los métodos basados en una descripción volumétrica de las formas generalmente dan buenos resultados aún para formas complejas con topologías diferentes, por lo que cualquier tipo de interpolación continua producirá una transformación suave [44].

Metamorfosis de Volúmenes con Transformada de Fourier. El método presentado por John F. Hughes [48] tiene como objetivo transformar un modelo volumétrico a otro, de forma automática y en base a señales. Se considera una transformación de Fourier sobre las funciones que definen los objetos, induciendo una transición entre superficies iguales de los dos modelos. La técnica se basa en interpolar suavemente entre las transformadas de Fourier de los dos modelos volumétricos y después transformar los resultados de regreso. En esta técnica, se trata de remover las características del modelo y solamente interpolar la forma general, o en su lugar, agregar aleatoriamente las características añadiendo ruido de alta frecuencia a la información; después interpolar los volúmenes subyacentes y finalmente, reducir el ruido de alta frecuencia. En este caso, la metamorfosis entre los detalles de los modelos se oculta mediante ruido.

Metamorfosis de Volúmenes basada en Características. Lerios et al. [49] presentan una técnica que es una extensión del trabajo de Beier y Neely [41], pero a diferencia de éste, se trabaja con objetos 3D y el usuario especifica las características correspondientes ya no sólo por medio de líneas, sino que también puede usar puntos, rectángulos y cajas.

Con el propósito de obtener una buena metamorfosis, es necesario especificar una colección de pares de características que definan la correspondencia global de los dos objetos. Cada par

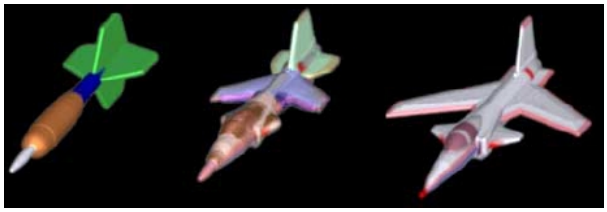

Figura 2-13: Metamorfosis basada en características para la representación volumétrica de un dardo y un avión.

de características define un campo que se extiende por todo el volumen. Una colección de pares de características define una colección de campos, en el que todos influencian a cada punto en el volumen. Entonces se usa un esquema de promedio de pesos para determinar el punto en el objeto fuente que corresponde a cada punto del objeto destino (Figura 2-13).

Metamorfosis 3D basada en Campos de Distancia. La técnica de Daniel Cohen-Or et al. [43] es para una metamorfosis 3D basada en interpolación de campos de distancia. Al principio los objetos fuente y destino se pasan a una representación de volúmenes discretos de campo de distancia. El método consta de dos pasos: warping e interpolación. El warping se usa para deformar el espacio 3D para hacer que los dos objetos coincidan tanto como sea posible. La interpolación es lineal y se aplica a campos de distancia deformados por el warping. La interpolación de campo de distancia (Distance Field Interpolation - DFI) es una interpolación general de conjuntos valuados para la reconstrucción de un modelo n-dimensional a partir de una secuencia de sus intersecciones $(n-1)$-dimensionales.

El método de DFI logra una mejor reconstrucción de la superficie al interpolar los valores de distancia. Trabaja bien debido a que las partes correspondientes de los dos objetos se alinean apropiadamente. De otra forma, algunas partes pueden desaparecer y reaparecer inesperadamente. Para superar esta restricción se combina DFI con una transformación de distorsión apropiada.

Después de revisar las técnicas que se utilizan en la animación facial y en particular la técnica de metamorfosis, es tiempo de pasar a otro tema esencial para la sincronización del movimiento de labios: la relación entre sílabas, fonemas y visemas que se expone en el siguiente capítulo.

Capítulo 3

Relación Sílaba - Fonema - Visema

Tradicionalmente, cualquier sistema de animación de movimiento de labios manejado por texto o voz usa fonemas como las unidades básicas del habla y visemas como las unidades básicas de la animación. El término *visema* viene del inglés *viseme* y fue acuñado inicialmente por Fisher [50] como una amalgama de las palabras *visual* y *phoneme*. A la fecha, no hay una definición precisa del término, pero en general hace referencia a un segmento del habla que es *visualmente* contrastante de otro [35].

En este capítulo se presenta un breve estudio de fonología del idioma español en base al trabajo de Núñez Cedeño y Morales-Front [51], empezando por analizar los órganos del habla y su intervención en la articulación de sonidos. Posteriormente, se discute el proceso de silabificación en español, para presentar la relación entre sílabas, fonemas y visemas.

3.1. Los Órganos del Habla

Es conveniente pasar revista a los órganos que participan en la articulación de los sonidos del habla. Esto nos permitirá tener una idea más precisa de la configuración del aparato fonador, según se muestra en la Figura 3-1 y del estado de cada órgano en la articulación de cada sonido.

Es importante tener presente que los órganos que intervienen en la producción de los sonidos del habla tienen como función primordial el facilitar la respiración y la alimentación. Es decir, los pulmones y los dientes no aparecieron en los humanos bajo el imperativo de facilitar la comunicación, sino para facilitar la respiración y el procesamiento inicial de los alimentos

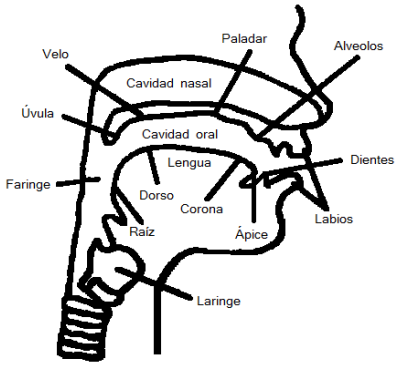

Figura 3-1: Los órganos del habla.

respectivamente. Más allá de estas funciones básicas, los humanos hemos superpuesto en esos mismos órganos la función de articular sonidos. La mayor parte de los articuladores se pueden observar desde el exterior y por ello son fáciles de identificar. Los labios, dientes y alveolos no necesitan ser descritos. El cielo de la boca se divide en dos secciones, el paladar es la parte dura y el velo la parte blanda. Al final del velo está la úvula (la campanilla) que controla el acceso a la cavidad de resonancia nasal. El ápice es la parte superior de la lengua (la punta). La lámina, o corona, comprende una pequeña área justo detrás del ápice. El dorso de la lengua empieza donde termina la corona y se extiende hasta la altura de la úvula. Finalmente, la raíz está en la base, más allá de la úvula.

Entre estos órganos, algunos son activos en el sentido de que inician el movimiento articulatorio, mientras que otros son pasivos y simplemente reciben el contacto del articulador activo. El articulador activo por excelencia es la lengua (especialmente el ápice y la corona), pero también son activos el labio inferior, la úvula y la laringe. Los articuladores pasivos son: el labio y dientes superiores, los alveolos, el paladar, el velo y la faringe. Los articuladores activos normalmente buscan el articulador pasivo más cercano: así, el labio inferior se desplaza hacia el superior, el ápice hacia los dientes o los alveolos, la corona hacia el paladar, el dorso hacia el velo o la úvula y la raíz hacia la faringe. Sin embargo, hay casos en los que un articulador activo se desplaza más allá de su dominio habitual. Esto pasa, por ejemplo, con las consonantes retroflejas que se articulan con el ápice de la lengua marcadamente curvado hacia atrás, de forma que alcance el paladar.

3.2. Clasificación de los Sonidos

3.2.1. Según las Posibilidades de la Laringe

La laringe es un órgano de particular importancia para el habla porque en ellas residen las cuerdas vocales. Las cuerdas vocales son los músculos que presentan un primer obstáculo al aire procedente de los pulmones. Pueden vibrar si la tensión de los músculos y la presión del aire están dentro de unos límites determinados.

En un sonido sordo las cuerdas vocales se mantienen separadas al grado de que no haya ningún tipo de resistencia al paso del aire por la glotis. Ese es el caso en la pronunciación de sonidos como [s] en el primer sonido de *sé*, [θ] como en el primer sonido de *zinc*, [f] *fé*, [č] *Ché*, [x] *José*, [t] *té*, [p] *pan*, o [k] *casa*. Si la tensión de las cuerdas es mayor de lo que consideramos una posición relajada y en consecuencia, se produce una cierta turbulencia al paso del aire, se trata de una articulación fricativa glotálica y el sonido producido se transcribe con el símbolo [h]. Este es el sonido que los dialectos aspirantes del español pronuncian en ciertos contextos en lugar de la /s/[1]. Es también el primer sonido en la pronunciación inglesa de la palabra *home* 'casa'.

En un sonido sonoro las cuerdas vocales se acercan lo suficiente como para producir un momentáneo aumento de la presión en la zona subglotálica. Cuando el aumento de la presión basta para desbordar la resistencia que ofrecen las cuerdas vocales, el aire se libera y como resultado la presión detrás de las cuerdas baja repentinamente. Al reducirse la presión, la tensión de las cuerdas vocales es de nuevo suficiente para obturar el paso del aire, con lo que se reinicia el ciclo de presión y descompresión. Ese continuo abrir y cerrar del paso del aire en la zona glotal es lo que genera la vibración característica de los sonidos sonoros. La vibración de las cuerdas en estos sonidos es fácilmente perceptible si se mantienen los dedos contra la glotis (la nuez) mientras se articula una vocal como en el primer sonido [a] en la palabra *ana*, [R] como en *rana*, [l] en *lana*, [n] *nana*, [m] *mona*, [b] *vana*, [g] *gana*, [d] *daga*, o [w] como en *hueso*.

Sin embargo, las posibilidades de las cuerdas vocales no se limitan a contrastar sonidos sonoros y sordos. Las cuerdas vocales son también las que controlan el susurro, una modalidad

[1] La mayoría de los sistemas notacionales suelen distinguir entre sonidos (entre corchetes) y fonemas (entre barras oblicuas).

de fonación en la que se reduce la resonancia de todos los sonidos de modo que resulta ideal para comunicar secretos al oído. La voz suspirante es un tipo de fonación que prolonga la articulación de algunas vocales con un suspiro. Por otra parte, la voz vibrante es un tipo de fonación que se caracteriza por una vibración de las cuerdas de frecuencia muy lenta. En el español, estos tipos de fonación no se usan distintivamente y por eso los concebimos como efectos especiales de la voz. Hay que tener presente, sin embargo, que en otras lenguas, estos "efectos especiales" funcionan contrastivamente. Es decir, crean distinciones equivalentes al contraste entre sonidos sordos y sonoros o al de /r/ frente a /l/.

3.2.2. Según el Grado de Constricción en la Cavidad Oral

Según el grado de constricción, el tipo de contraste más utilizado por todas las lenguas es el que existe entre vocales y consonantes. Las vocales son los sonidos que, debido al grado de apertura de la cavidad oral, tienen mayor sonoridad o perceptibilidad. Una consonante puede definirse como un sonido cuyo grado de obstrucción es mayor al de cualquier vocal.

Vocales

Los sonidos vocálicos se generan con una vibración inicial de las cuerdas vocálicas y la posterior modificación de estas vibraciones en las cavidades supraglotálicas. La configuración de esas cavidades puede modificarse según la posición de la lengua y los labios. En la medida en la que el tamaño de la cavidad se altera, así cambia también para un oyente la percepción del sonido emitido. Por tanto, la diferencia entre /i/ y /e/ resulta de la diferencia en el tamaño de la cavidad oral. En el caso de la /e/ la cavidad es ligeramente mayor y esto tiene un efecto de resonancia claramente perceptible.

En principio, cada vocal podría ir acompañada o no de una protuberancia y redondeamiento de los labios. En la práctica, sin embargo, hay muchos sistemas en los que la protusión labial es característica sólo de las vocales posteriores. El hecho de que esta correspondencia entre posterioridad y redondeamiento no tenga excepciones en español es un indicativo de que la protusión labial de las vocales no funciona contrastivamente dentro del sistema fonológico.

Otra característica fonética que puede dar lugar a la duplicación de elementos dentro del espacio vocálico es la duración relativa. La duración relativa ignora las diferencias de articulación

inherentes a cada vocal. Por ejemplo, las vocales bajas suelen tener una mayor duración debido al mayor esfuerzo articulatorio que supone el desplazamiento de la mandíbula para crear la apertura necesaria en la cavidad oral. De forma paralela, la duración de las vocales puede variar dependiendo de la consonante que sigue. Por ejemplo, en inglés, las vocales son consistentemente más largas cuando van seguidas de una consonante sonora que seguidas de una sorda. Al margen de estas variaciones intrínsecas o contextuales de duración, hay lenguas en las que una [a] que dure aproximadamente el doble que una [a] simple se percibe como una vocal diferente. Ese era el caso en latín clásico, por ejemplo, cuyo sistema constrastaba sistemáticamente entre vocales largas y breves. En todas las variedades del español moderno este contraste se ha perdido aunque quedan algunos vestigios que pueden rastrearse en la diptongación de las vocales medias.

Las vocales del español se caracterizan por su uniformidad y nitidez. La nitidez se debe al hecho de que se articulan en la periferia del espacio vocálico. La uniformidad se debe al hecho de que el hablante mantiene la posición de los articuladores a lo largo de toda la duración de la vocal. Esto contrasta con lenguas como el inglés, cuyas vocales se deslizan hacia constricciones más altas durante su articulación. Otra característica que contribuye a la robustez del sistema vocálico del español es el hecho de que no haya una tendencia a la reducción de las vocales cuando no llevan acento prosódico. Con excepción de la diptongación de algunas vocales medias bajo la influencia del acento, la pronunciación de las vocales tónicas y átonas sólo varía con respecto a su duración media.

Consonantes

Cualquier constricción en la cavidad oral que tenga un grado superior al de la vocal /i/ es por definición una consonante.

Grado de Constricción. Empezaremos por ver cómo se clasifican las consonantes según el grado de aproximación del articulador activo al pasivo.

Aproximantes. Estas son las consonantes con el grado de constricción más cercano al de una vocal, lo que las caracteriza es que no dificultan el paso del aire lo bastante como para que se cree una turbulencia o fricción perceptible. En este grupo entran la /r/ de *pera* y la /l/ de

lejos. En estos sonidos el aire pasa por la cavidad oral sin que se perciba una fricción. Las aproximantes, como las vocales, son casi siempre sonoras. Es decir, van normalmente acompañadas de vibración en las cuerdas vocales. La explicación de esta dependencia resulta transparente cuando consideramos cómo se articulan estos sonidos. Al igual que los sonidos vocálicos, los sonidos aproximantes necesitan una vibración básica cuya resonancia pueda variar según el tamaño de la cavidad oral. Sin la vibración básica las diferencias de tamaño no son lo bastante perceptibles en sí mismas, ya que estos sonidos no crean turbulencia en la cavidad oral. Por supuesto, existen vocales, nasales y aproximantes sordas, pero cierto es que estos sonidos suelen tener una distribución muy limitada. La perceptibilidad de estos sonidos sordos radica en el hecho de que la falta de vibración en las cuerdas vocales se compensa con algún tipo de turbulencia en el área glotálica. Eso es similar a lo que sucede, por ejemplo, en el susurro. Las cuerdas no tienen una tensión suficiente para que se produzca una vibración pero están lo bastante juntas como para que el aire no pase libremente. Esto basta para crear una onda básica cuya resonancia puede manipularse en la cavidad oral.

Las aproximantes pueden ser centrales o laterales. En las centrales, la lengua está en contacto con la parte interior de los dientes a la altura de los primeros molares y sólo se aproxima al paladar, dejando escapar el aire por el centro. En las laterales la zona de contacto y aproximación se invierte. Hay contacto por el centro en los alveolos o en el paladar pero la lengua sólo se aproxima a la parte interior de los molares dejando escapar el aire por el lado.

Fricativas. Si el grado de constricción se lleva al punto en que empieza a formarse una turbulencia en el aire detrás del punto en que el articulador activo y pasivo se acercan uno a otro, se produce un sonido fricativo. La calidad fricativa de sonidos como la /f/ en *falso*, /θ/ en *zona*, /s/ en *sol*, y /x/ en *jaca*, es claramente distintiva y se usa contrastivamente en español. La fricativa glotal [h] es muy común en los dialectos aspirantes.

A diferencia de lo que pasa con sonidos de mayor apertura, en cuya perceptibilidad tiene un papel fundamental la vibración de las cuerdas vocales, las fricativas pueden ser tanto sonoras como sordas. De hecho las fricativas sordas son más comunes que las sonoras. Esto se debe a que la fricción que se produce al acercar los articuladores es perceptible en sí misma. En estos sonidos, la finalidad de la constricción es más la de crear una fricción que la de crear una

cámara de resonancia de un tamaño específico. La vibración glotal es perceptible cuando se combina con la fricción supraglotal, sin embargo no añade nada a la perceptibilidad inherente de la consonante fricativa.

Oclusivas. Son las consonantes con un grado máximo de constricción oral. Puesto que la aproximación de los articuladores es total y que el velo se mantiene subido cerrando el paso del aire a la cavidad nasal, se produce un bloqueo temporal en la salida del aire. Como el impulso de los pulmones no se detiene durante este bloqueo, el resultado es un aumento de la presión detrás del punto donde los articuladores entran en contacto. Lo que se percibe en estos sonidos es la ausencia momentánea de sonido y la súbita explosión que se produce al liberar el aire atrapado. Igual que con las fricativas, el objetivo en estos sonidos no es el de producir resonancias sobre la base de una vibración o turbulencia previa en el área de la glotis. Son perceptibles aunque las cuerdas vocales estén completamente relajadas. Las oclusivas también pueden ser sordas o sonoras, pero hay que señalar que la perceptibilidad de la vibración de las cuerdas vocales se ve atenuada como resultado de la oclusión momentánea en la cavidad oral. Esa incompatibilidad de las oclusivas con la sonoridad puede verse reflejada en procesos de alargamiento fonético compensatorio de las vocales ante oclusivas sonoras, como en inglés o en la fricativización de las oclusivas sonoras como en español. Las consonantes oclusivas del español son /p/ como en *palo*, /t/ en *todo*, /k/ en *cada*, /b/ en *bala*, /d/ en *dama* y /g/ en *gamo*.

Nasales. Al igual que en las oclusivas, en las consonantes nasales hay una total obstrucción de la cavidad oral. Sin embargo, en las consonantes nasales la úvula desciende separándose de la pared faríngea y esto crea una apertura que permite que el aire escape por la cavidad nasal. La cavidad nasal, al estar llena de membranas, actúa como una sordina que amortigua el sonido, dándole así la calidad típicamente nasal que escuchamos en /m/ *mano*, /n/ *nana*, /ñ/ *caña*. Si la úvula se baja pero no hay una oclusión en la cavidad oral se produce una vocal nasal como las del portugués, o el francés. En español las vocales que preceden a las consonantes nasales pueden pronunciarse con cierta nasalización.

Africadas. Las africadas pueden verse como una combinación de una oclusiva y una fricativa. Su articulación empieza con un cierre total del conducto oral, como en las oclusivas,

pero la apertura no es súbita sino que la articulación se desliza a un grado de constricción ligeramente inferior que al mantenerse produce una fricción. En el español moderno las únicas africadas son [č] y [ŷ] como en *noche* y *yema*. En otras lenguas, sin embargo, o incluso en etapas previas del español, las posibilidades de las africadas son, o eran, mayores.

La R mútiple. La vibrante múltiple es esencialmente similar a la vibrante simple con la diferencia de que suele tener más de una vibración. En la pronunciación de este sonido, la corona se aproxima a los alveolos y se pone en tensión. El ápice se mantiene con una tensión menor que ofrezca una leve resistencia al paso del aire. Esta resistencia hace que la presión detrás de la lengua aumente. Cuando la presión aumenta lo suficiente, el ápice se desplaza hacia abajo, pero no la corona (dado que esta parte de la lengua está más tensa). El aire puede entonces escapar y como consecuencia la presión detrás de la lengua se reduce. Cuando la fuerza del ápice es de nuevo superior a la presión, el ápice vuelve a cerrar el paso del aire y se inicia un nuevo ciclo. La sucesión rápida de estos ciclos crea la vibración múltiple característica de la [R] en la palabra *Ramón*.

La distribución de la [r] y la [R] en español es complementaria en todos los contextos excepto entre vocales, donde estos dos sonidos parecen contrastar como lo demuestran pares mínimos del tipo *corro*/*coro*. Esto hace que debamos plantearnos si se trata de dos fonemas contrastivos o de alófonos de un mismo fonema. Por una parte estos dos sonidos dan lugar a pares mínimos como los fonemas, pero por otra, en la mayoría de los contextos tienen una distribución complementaria como los alófonos. La posición más aceptada es la de dar por sentado que hay un sólo fonema subyacente y que las diferentes manifestaciones contextuales se derivan a base de reglas, mientras que el contraste intervocálico resulta de preespecificar una unidad extra de tiempo.

Punto de articulación. Hasta aquí se han analizado distintos modos de articulación según su grado de constricción. Sin embargo, el mismo grado de constricción se percibirá como un sonido distinto cuando se articule con el ápice contra los dientes o con el dorso contra el velo. Por ejemplo, si consideramos una oclusión, en el primer caso tendremos /t/ y el segundo /k/, o si se trata de una fricativa /θ/ y /x/. El contraste entre estos pares es de punto de articulación. El punto de articulación se determina con respecto al articulador pasivo, al que el articulador

activo toca o se aproxima. Según su punto de articulación, las consonantes se clasifican del modo siguiente:

Labiales. Son sonidos articulados contra el labio superior, /p/ *Pedro*, /b/ *bebé*, /m/ *mamá*, tienen este punto de articulación.

Labiodentales. Se articulan con el labio inferior contra los dientes superiores. La labiodental por excelencia en español es /f/, pero por asimilación de una nasal a /f/ puede también producirse una labiodental, nasal, fricativa. Una ausencia destacable es la de la labiodental, fricativa, sonora /v/, pero hay que notar que esto no es una particularidad de las labiodentales sino que puede extenderse a todas las fricativas sonoras en todos los puntos de articulación.

Interdentales. Se articulan colocando el ápice entre los incisivos superiores e inferiores. Quizás debido a que los dientes pueden dificultar la oclusión en caso de un alineamiento muy imperfecto, o por poder perderse con la edad o debido a accidentes, lo cierto es que las oclusivas interdentales, al igual que las labiodentales, no son nada comunes. Los sonidos consonánticos que se articulan contra la punta de los incisivos son exclusivamente fricativos. El sonido interdental más conocido del español es [θ], pero hay que notar que en la mayoría de los dialectos del español este sonido se ha perdido al confundirse con [s] en una de las finales estribaciones en la reestructuración del sistema de sibilantes medievales.

Dentales. Se articulan con el ápice contra la pared posterior de los dientes. En español tienen ese punto de articulación /t/ *todo* y /d/ *dar*.

Alveolares. Las alveolares se producen con el ápice o la corona contra los alveolos. Tenemos /s/ *sal*, /n/ *no*, /l/ *lazo*, /r/ *mar* y /R/ *rosa*. No hay oclusivas alveolares y esto debe atribuirse al hecho de que hay oclusivas dentales. Debido a su cercanía no se suelen contrastar consonantes con un mismo grado de constricción en esos dos puntos de articulación. La ausencia de la fricativa sonora /z/ está en consonancia con la ya observada ausencia de fricativas sonoras en general, pero como en casos anteriores este sonido se da alofónicamente como resultado de la sonorización de /s/.

Retroflejas. Se articulan a base de curvar la lengua hacia atrás hasta tocar con el ápice el paladar duro. El español no usa contrastivamente consonantes retroflejas pero hay dialectos en los que abundan estos sonidos.

Palatales. Se articulan con el dorso de la lengua contra el paladar. Las palatales del español son /č/ *chico*, /ñ/ *baño* y /y/ *calle*. La distribución de las palatales tiene limitaciones posicionales que derivan mayoritariamente del hecho de que desde un punto de vista diacrónico son de incorporación relativamente reciente al inventario.

Velares. En su articulación intervienen el dorso de la lengua y el velo del paladar. Las oclusivas velares son /k/ *canasta* y /g/ *goma*. La fricativa velar es /x/ como en *jarro*. La pronunciación de la fricativa /x/ admite mucha variación dialectal en punto y modo de articulación.

Uvulares. Se articulan con el dorso de la lengua contra la úvula. El español no usa contrastivamente sonidos producidos en esa área pero una vez más tenemos que notar la presencia de sonidos uvulares en dialectos españoles.

Faríngeas. Se articulan con la raíz de la lengua contra la pared faríngea. No se usan contrastivamente en español, pero hay pronunciaciones de la [χ] castellana que se aproxima a una pronunciación faríngea.

3.3. La Silabificación en Español

3.3.1. La Sílaba

Fonológicamente los segmentos (consonantes y vocales) se agrupan en sílabas. La sílaba se puede definir como un conjunto de segmentos agrupados en torno a un núcleo (la vocal). En español suele haber acuerdo entre los hablantes no sólo respecto al número de sílabas, sino también sobre la posición exacta del límite entre dos sílabas. Por esto no suele haber ninguna indecisión al dividir una palabra en la frontera entre dos sílabas al final de una línea. En otras lenguas, como el inglés, sin embargo, la división en sílabas resulta más problemática.

Dentro de la sílaba, los segmentos se organizan de acuerdo con una escala universal de sonoridad (Figura 3-2a), de tal modo que el segmento con mayor sonoridad ocupa el lugar central en la sílaba (el núcleo) y otros segmentos a su izquierda o su derecha han de descender progresivamente en sonoridad.

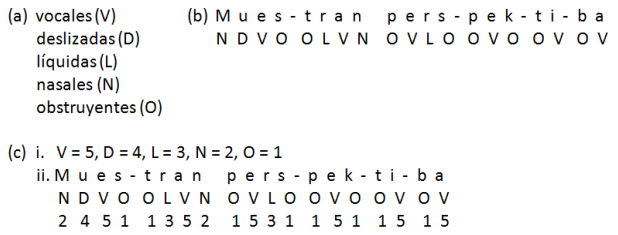

Figura 3-2: a) Escala de sonoridad. b) Ejemplo de distribución de sonoridad dentro de la sílaba. c) El valor de los segmentos en la escala disminuye progresivamente conforme se alejan del núcleo.

Los ejemplos en la Figura 3-2b muestran esta distribución dentro de la sílaba. Si asignamos un número a los diferentes segmentos según su posición en la escala de sonoridad, podemos observar cómo el valor de los segmentos en esta escala disminuye progresivamente según nos alejamos del núcleo (Figura 3-2c).

Así, pues, la sílaba consiste en un conjunto de segmentos agrupados alrededor de una cumbre o pico de sonoridad, de tal modo que los segmentos más cercanos al núcleo tienen un índice de sonoridad que nunca es menor que el de los más alejados. Debemos, sin embargo, notar que dos segmentos contiguos en la sílaba pueden tener el mismo índice de sonoridad, aunque esto raramente ocurra en español. Esto lo vemos en una secuencia de dos obstruyentes como la que encontramos en la segunda sílaba de *bíceps*. Dos segmentos de máxima sonoridad (dos vocales) también pueden ser adyacentes, constituyendo sílabas diferentes, como en *poeta*.

Para referirse a posiciones o grupos de segmentos dentro de la sílaba se utilizan los términos núcleo, ataque (o arranque), coda y rima. El *núcleo*, como hemos indicado, es el centro de la sílaba y el elemento de mayor sonoridad dentro de ella. En español todas las vocales constituyen núcleos silábicos y todos los núcleos silábicos contienen una vocal. Tal identidad no se da en lenguas como el inglés, que tienen consonantes silábicas. La única dificultad en establecer el

número de sílabas en español la encontramos en la silabificación de secuencias de vocoides[2] como *ai*, *iu*, *eu*, en las que los dos vocoides pueden realizarse como vocales, en dos sílabas separadas, [a-i], [i-u], [e-u], o uno de ellos puede ser una deslizada y la secuencia pertenecer por tanto a una sola sílaba, [aj], [ju], [ew].

El *ataque* es la consonante o grupo de consonantes que preceden al núcleo dentro de las sílabas. En español los únicos grupos de ataque permitidos son los que consisten en una oclusiva o /f/ seguida de líquida (como en *primo*, *grupo*, *broma*, *blanco*, *claro*, *flecha*, *freno*), excepto /dl/ y, según el dialecto, /tl/. Entre los dos miembros de un grupo de arranque ha de haber una cierta distancia en sonoridad, lo que excluye grupos de obstruyente más nasal como /pn/ que sólo difieren en un grado en la escala dada en la Figura 3-2a. Además de esto, no puede haber una semejanza excesiva en la articulación de los dos segmentos. Es evidente que lo que imposibilita tener /dl/ como grupo tautosilábico es la coincidencia en una serie de rasgos fonológicos entre los dos segmentos: ambos son coronales, sonoros y no-continuantes. En /tl/ la coincidencia es menor dado que el primer segmento es sordo y el segundo sonoro y esto se refleja en la mayor aceptabilidad del grupo (que es perfectamente posible en el español de México, aunque resulte excluido en el de Madrid). Una palabra como *atlas* se pronuncia [á-tlas] en casi toda Latinoamérica y áreas del oeste peninsular, mientras que en el centro y este de la península (ibérica) se pronuncia [át-las]. En español mexicano el grupo /tl/ ocurre incluso en principio de palabra, en topónimos y préstamos del nahuatl como *Tlaxcala*, *tlapalería*, etc.

La *coda* es lo que sigue al núcleo en la sílaba. Tras el núcleo podemos encontrar una deslizada como en *boi-na, flau-ta* o una consonante como en *pren-sa*, *ac-to*. Podemos tener también una deslizada seguida por consonante, como en *claus-tro*, *vein-te*, o un grupo de dos consonantes, como en *trans-por-te*, *pers-pec-ti-va*. No todas las consonantes son igualmente comunes en posición de coda. Debemos distinguir entre codas finales y codas interiores. En final de palabra podemos encontrar en principio sólo consonantes coronales: las resonantes /-l/, /-n/, /-r/ (*papel*, *camión*, *amor*), y las obstruyentes /-s/, /-θ/ (en dialectos con este fonema, *mes*, *pez*) y /d/ (*virtud*). Excepcionalmente podemos tener /-x/ (*reloj*) y /-t/ (*cénit*). Otras consonantes finales aparecen únicamente en préstamos no completamente asimilados como *club*, *frac*, *bulldog*, *álbum*, *chef*. En cuanto a las consonantes que son posibles en coda interior de palabra, podemos tener una

[2] Vocales y deslizadas forman la clase de los vocoides.

nasal con el mismo punto de articulación que la consonante siguiente, como en la primera sílaba de *campo*, *canto* o *tengo*, un líquida /l/, /r/, como en *palco*, *parco* (neutralizadas en dialectos caribeños y andaluces) o una sibilante /s/ o /θ/ (en dialectos con esta distinción), como en *asco*, *juzga*. Sin embargo, en el habla no suelen hacerse tantas distinciones como las indicadas ortográficamente ni siquiera en pronunciación cuidada. Aunque se pronuncie la obstruyente postvocálica es común que no haya distinción alguna entre sordas y sonoras en esta posición. Lo normal es que la *p* de, por ejemplo, *concepción* y la *b* de *obsesión* tengan idéntica realización fonética y esto es también cierto para la *t* y la *d* de *étnico* y *administrar*, por ejemplo, o la *c* y la *g* de *técnica* y *dogmático*.

Las codas de dos segmentos son poco frecuentes. En estos casos podemos encontrar una deslizada seguida por una consonante (*au[k-s]ilio*, *vei[n-t]e*, *au[n-k]e*) o una consonante seguida de /s/ (*abstracto*, *bíceps*, *vals*, *adscrito*, *transporte*, *experiencia*, *perspectiva*). En el léxico común las codas de dos consonantes tienen siempre /s/ como segundo elemento.

Finalmente, núcleo y coda se agrupan en una unidad superior denominada *rima*. Los términos ataque, núcleo. coda y rima definen, pues, estructuras como la ilustrada en la Figura 3-3.

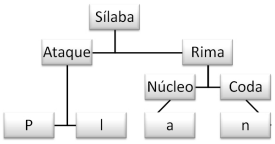

Figura 3-3: Representación de la sílaba *plan*.

3.3.2. Principales Generalizaciones acerca de la Silabificación en Español

El español es una lengua de estructura silábica relativamente sencilla. Cuando consideramos cómo se distribuyen en sílabas los segmentos en español, una primera observación que podemos hacer es que una consonante intervocálica se silabifica siempre con la vocal siguiente. Así *copa* es [ko-pa] y no [kop-a], aunque esta segunda silabificación produciría también sílabas que, tomadas aisladamente, estarían bien formadas, como se ve si comparamos las sílabas obtenidas con la

primera de las palabras [kop-to] y [a-to]. El principio de que una consonante se silabifica siempre con la vocal siguiente se aplica incluso cuando entre vocal y consonante media un límite de palabra, como en *las alas* [la-sa-las] (igual a *la salas* 'le pones sal'). Esta silabificación a través de fronteras morfológicas o sintácticas no ocurre sin embargo en casos como *un hueso* [uŋ-gwé-so] o *deshielo* [dez-ŷé-lo] porque las deslizadas en posición inicial de palabra (o raíz) se realizan como consonantes. No tenemos, pues una secuencia C-V[3] en estos ejemplos, sino una secuencia de consonantes.

Otra generalización que podemos observar es que en la silabificación de grupos de consonantes, aquellas secuencias que pueden dar lugar a grupos de ataque legítimos (los constituidos por oclusiva o /f/ seguida de líquida, excepto /dl/ y dialectalmente /tl/) se silabifican como ataque, aunque, de nuevo, otras silabificaciones produjeran estructuras bien formadas desde el punto de vista de los tipos de sílaba permitidos en español. Así, por ejemplo, tenemos *soplo* [so-plo] y no [sop-lo]. Nos referimos a este fenómeno como maximización de ataques silábicos. Sin embargo, al contrario de la generalización anterior, la maximización de ataques silábicos no es un principio que se aplique entre palabras, *chef loco* no puede ser [če-flo-ko] sino [čef-lo-ko] y *club latino* no puede silabificarse como [klu-βla-ti-no] sino que la única silabificación posible es [kluβ-la-ti-no] donde las fronteras silábicas coinciden con los límites de palabra. Los mismos efectos se observan incluso en interior de palabra con ciertos prefijos, como en *sublingual* [suβ-liŋ-gwal] (no [su-βliŋ-gwal]), donde el grupo /bl/ que puede formar ataque silábico, aparece sin embargo dividido por una frontera silábica que corresponde a la morfológica.

La maximización de ataques silábicos se obtiene ordenando la formación de ataques antes de la formación de codas. Veamos un análisis posible de la silabificación en español utilizando reglas ordenadas.

El primer paso ha de ser la identificación de los núcleos silábicos (Figura 3-4). El núcleo es el único elemento obligatorio de la sílaba. Puede haber sílabas sin arranque ni coda, como la primera sílaba de *ala*; pero no sílabas sin núcleo. En español encontramos el caso, bastante sencillo, de que los únicos segmentos que actúan como núcleos de sílaba son las vocales. Esto es, al contrario que en otras lenguas, en español una consonante no puede constituir el núcleo de una sílaba.

[3] C = consonante, V = vocal

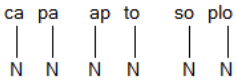

Figura 3-4: Identificación de núcleos silábicos. El núcleo se indica como N.

La siguiente operación es la aplicación de la llamada regla CV, que silabifica una consonante como ataque (bajo el nodo A) con una vocal inmediatamente a su derecha (Figura 3-5). La aplicación de la regla CV como primera regla de adjunción produce el resultado de que en una secuencia VCV la consonante intervocálica se agrupe con la vocal siguiente y no con la precedente (es decir, [os-o] es imposible de generar).

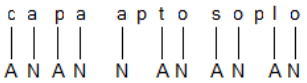

Figura 3-5: Regla CV (incorporación de consonantes prevocálicas). El ataque se identifica por A.

La maximización de ataques silábicos (i.e. [ko-pla] y no [kop-la]) se consigue aplicando acto seguido una segunda regla de adjunción que silabifica una segunda consonante en el ataque siempre que el resultado sea un grupo de ataque admisible. Entre los ejemplos de la Figura 3-6, esta regla sólo tienen aplicación en *soplo*, adjuntando, la /p/al ataque de la segunda sílaba. En *apto*, sin embargo, la regla de formación de grupos de ataque no tiene efecto porque su aplicación crearía un grupo /pt/, que no constituiría un ataque aceptable en español. La aplicación de esta segunda regla de adjunción bajo el nodo A requiere, pues, la definición independiente de los grupos de ataque admisibles en la lengua (oclusiva o /f/ más líquida, excepto /dl/ y dialectalmente /tl/).

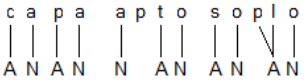

Figura 3-6: Formación de grupos de ataque.

Sólo una vez que estas dos operaciones de adjunción en posición de ataque se hayan aplicado,

se aplicarán otras reglas, adjuntando otros segmentos en posición de coda (bajo el nodo C; se añade también el nodo R -rima- a la representación en la Figura 3-7).

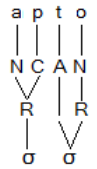

Figura 3-7: Adjunción de codas. La coda se representa por C, la rima por R, y la sílaba entera por σ.

Dado este orden de reglas, silabificaciones incorrectas como [kap-a] y [sop-lo] son imposibles de generar. Para que esto sea así, resulta crucial aplicar las reglas en el orden indicado. Un cambio en el orden de operaciones produciría resultados no deseados. De modo que (ignorando por el momento estructuras más complejas que nos llevarían a postular algunas reglas adicionales), hemos reconocido cuatro operaciones ordenadas de silabificación:

- Identificación de núcleos.
- Regla CV.
- Regla de Formación de Grupos de Ataque.
- Reglas de Adjunción de Codas.

3.3.3. Hiatos y Diptongos

Hasta ahora hemos omitido casi completamente toda discusión sobre la silabificación de las deslizadas y más generalmente sobre diptongos e hiatos. Esto es porque estas secuencias presentan problemas especiales en español, que examinaremos en esta sección.

Para dar cuenta de la silabificación de secuencias vocálicas necesitamos modificar la escala de sonoridad dada en la Figura 3-2(a) asignando valores diferentes a distintas vocales según su grado de apertura, como se muestra en la Figura 3-8.

Dos vocoides seguidos pueden pronunciarse juntos en una sola sílaba (formando diptongo), como *io* en *Mario*, o en sílabas separadas (en hiato) como *ia* en *María*. En español tenemos tanto

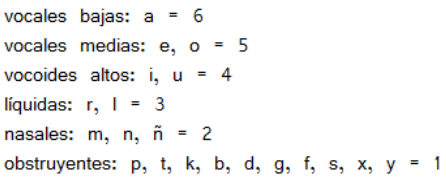

Figura 3-8: Escala de sonoridad con valores diferentes para distintas vocales según su grado de apertura.

diptongos de sonoridad creciente, [ia], [ue], como diptongos decrecientes, [ai], [eu]. Asimismo podemos tener hiatos de sonoridad creciente (como [i-a], [u-e]) y de sonoridad decreciente (como [a-i], [e-u]). Cuando ninguna de las dos vocales en la secuencia es alta, tenemos siempre un hiato, al menos a nivel léxico o en pronunciación cuidada: *poeta* [po-é-ta], *maestro* [ma-és-tro], *teatro* [te-á-tro]. En el habla rápida, sin embargo, estas secuencias pueden reducirse también a diptongo. Las secuencias que, a nivel léxico, pueden formar diptongo en español son las que se ejemplifican en la Figura 3-9.

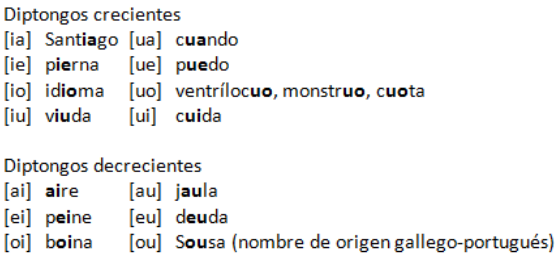

Figura 3-9: Los diptongos del español.

Observemos también que las secuencias *iu*, *ui*, de dos vocales altas pronunciadas en diptongo se han clasificado como diptongos crecientes, en vez de diptongos decrecientes. En realidad estas dos realizaciones son difícilmente distinguibles y parece haber preferencias diferentes en algunos dialectos. Lo general, sin embargo, es que *viuda* rime con *suda* y no con *vida*, lo que nos lleva a concluir que en el diptongo *iu* la vocal es [u], mientras que *cuida* rima con *vida*, lo que indica que en *ui* la vocal es [i].

Las mismas secuencias que pueden formar diptongo aparecen también en hiato en otras

palabras (Figura 3-10). Los hiatos se marcan ortográficamente con un acento cuando la vocal alta lleva el acento prosódico como en *María*, *navío*, *oído*, etc. Sin embargo, el acento no suele escribirse en 'pseudo-monosílabos' como *(él) rio* [Ri-ó] que en realidad es bisílabo y contrasta con los monosílabos *dio*, *vio*, donde el acento ortográfico no es necesario según las reglas de la Academia precisamente por tratarse de monosílabos. El acento tampoco se distingue ortográficamente del diptongo cuando las dos vocales son altas como en *huida*, *huimos*, que, sin embargo, tienen una secuencia con hiato que contrasta fonológicamente con el diptongo de *cuida*, *fuimos*. Esto es, *huida*, por ejemplo, tiene tres sílabas, exactamente como *oído*, mientras que *cuida* tiene sólo dos. Para muchos hablantes hay también un contraste entre, por dar otro ejemplo, *riendo*, con hiato y *siendo*, con diptongo, ambos con acento prosódico en la vocal [e] que tampoco se marca ortográficamente al no ser la vocal alta la que tiene acento prosódico.

[i-a]	María	[u-a]	púa
[i-e]	ríe	[u-e]	adecúe
[i-o]	navío	[u-o]	adecúo
[i-u]	diurno	[u-i]	huída
[a-i]	caída	[a-u]	aúlla
[e-i]	leímos	[e-u]	reúne
[o-i]	oímos		

Figura 3-10: Hiatos.

Las secuencias de sonoridad creciente con acento prosódico sobre la vocal no alta como /iá/, /ié/, /ió/, /uá/, etc., se silabifican generalmente como diptongos. Los casos con hiato donde la vocal alta no es la acentuada prosódicamente son la excepción y generalmente corresponden a palabras relacionadas morfológicamente con otras donde la vocal alta lleva el acento como en *riendo*, que pertenece al mismo verbo que *ríe*, donde la [i] lleva el acento, *riada* [Ri-á-da], relacionada con *río*, *viable* relacionada con *vía*. También encontramos hiatos en casos de prefijación como en *reúne* [Re-ú-ne] de *re + une*. Finalmente, para muchos hablantes, aunque al parecer no para todos, hay palabras que excepcionalmente tienen hiato sin que exista ninguna explicación morfológica para ello. Por ejemplo, mientras que *diente*, *mientras*, *vientre*, *siente*, *tiene* tienen un diptongo [ie], hay muchos hablantes para quienes la palabra *cliente* es diferente de las otras y contiene un hiato [kli-én-te]. De todo esto se concluye que en español existe un contraste fonémico entre vocales y deslizadas. De todas formas, cuando el acento no cae sobre la

vocal alta, las palabras con hiato son la excepción y encontramos variación dialectal en cuanto a qué palabras permiten la pronunciación en hiato.

3.4. Correspondencia entre Fonemas y Visemas

Generalmente, el *fonema* se ha definido como la unidad mínima de descripción fonológica, una entidad abstracta con una *función distintiva* a la que no corresponde una realidad fónica particular. Como función distintiva se entiende la capacidad de un sonido para producir una diferencia funcional o de distinción de significado en una lengua dada. La fonología parte generalmente del examen de los sonidos existentes en una lengua particular, para determinar qué unidades fonemáticas tienen una función lingüística. La *prueba de la conmutación*, propuesta por los fonólogos de la Escuela de Praga, consiste en probar la diferencia funcional entre fonemas a partir de su capacidad de generar distinciones de significado. Por ejemplo, la existencia de tripletes mínimos como *carro*, *sarro* y *tarro* en español demuestra que los sonidos [k, s, t] son fonemas distintos /k, s, t/ y capaces de generar una oposición distintiva [51]. Sin embargo, cuando se trata de la sincronización de movimiento de labios en la animación por computadora, basar esta sincronización en fonemas puede ser muy complicado.

En busca de un mejor sistema para la sincronización del movimiento de los labios, algo se volvió evidente: hay diferentes tipos de sonidos que se pueden hacer al hablar y no todos ellos son fáciles de ver. Como se expuso anteriormente en la clasificación de los sonidos, algunos son hechos principalmente por los labios, otros por la lengua y otros por la garganta y las cuerdas vocales. De éstos, de los únicos que se debe ocupar la animación es de los sonidos realizados principalmente por los labios.

Los fonemas son sonidos, pero lo que importa en la animación es lo que se puede ver. En lugar de usar fonemas, la sincronización del movimiento de labios se basa en *visemas*. Como se mencionó anteriormente, la palabra visema viene del inglés *viseme* que es una amalgama de las palabras *visual* y *phoneme*, es decir, fonema visual. Un visema puede definirse como una imagen estática de la forma de los labios que es *visualmente* contrastante de otra [35].

Debido a que la animación del movimiento de labios está basada en un conjunto de visemas, se necesita establecer una correspondencia entre el conjunto de fonemas de una lengua dada

y un conjunto de visemas que represente visualmente a estos fonemas. La aproximación más simple es asumir una proyección uno-a-uno entre el conjunto de fonemas y el conjunto de visemas. Actualmente, sin embargo, la investigación sobre fonemas indica que la proyección entre fonemas y visemas es muchos-a-uno: hay muchos fonemas que son similares visualmente y por lo tanto caen en la misma categoría de visema. Esto es particularmente cierto, por ejemplo, en casos donde dos sonidos son idénticos en la forma y punto de articulación, pero difieren sólo en características de voz. Por ejemplo, /b/ y /p/ son dos labiales que difieren solamente por el hecho de que la primera es sonora y la segunda sorda. Esta diferencia, sin embargo, no se manifiesta visualmente y por lo tanto los dos fonemas deben ser colocados en la misma categoría de visema.

Recíprocamente, la literatura apunta a que la proyección de fonemas a visemas también es uno-a-muchos: el mismo fonema puede tener diferentes formas visuales. Este fenómeno se conoce como *coarticulación* y ocurre debido a que el conexto fonémico en el cual un sonido es articulado, influencia la forma de los labios para ese sonido y suele constituir variantes fonéticas de los fonemas, o *alófonos*. Una prueba usada a menudo con el fin de identificar las posibles manifestaciones fonéticas de un fonema es el análisis de la variación de morfemas. Así, por ejemplo, se pueden observar varias realizaciones del morfema correspondiente al artículo indeterminado *un* en español; la nasal final adquiere distintos puntos de articulación, según cuál sea el punto de articulación de la consonante siguiente: u[n] árbol (no asimilada), u[m] burro (bilabial), u[*m*] fuego (labiodental), u[n̪] diente (dental), u[n] lomo (alveolar), u[ñ] yeso (alveolopalatal), u[ŋ] cuento (velar). A base de estos datos se puede deducir que /n/ es un fonema del castellano que se manifiesta fonéticamente en los siete alófonos anteriores.

Con respecto a la correspondencia entre visemas y fonemas, no existe una sola tabla o proyección en la literatura. Por lo general cada investigador utiliza una tabla diferente, aunque se pueden hacer algunas generalizaciones. Casi todos los autores coinciden en que el fonema /g/ no tiene un visema correspondiente (o por lo menos no uno fácilmente distinguible), ya que como se vio es una consonante velar, siendo su punto de articulación el dorso de la lengua y el velo del paladar. Entre las velares también se encuentran la /k/ y la /x/ como en *jarro*, por lo que también sus visemas correspondientes son difíciles de identificar.

A diferencia de estos fonemas, aquellos que tienen su punto de articulación en los labios o

los dientes son fácilmente relacionados con un conjunto de visemas que los representen. Este es el caso por ejemplo, de los fonemas /f/ y /v/ que se articulan con el labio inferior contra los dientes superiores. El proceso de proyección de fonemas a visemas es, en general, dependiente del idioma del que se trate, por lo que una proyección entre fonemas y visemas en inglés no provee una correspondencia entre fonemas y visemas en español. Como ejemplo, la Figura 3-11 muestra el conjunto de visemas obtenido por Tony Ezzat y Tomaso Poggio [35] para fonemas en inglés. Ellos grabaron a una persona enunciando un conjunto de palabras clave, después identificaron manualmente una sola imagen para cada visema y posteriormente redujeron el conjunto de visemas a dieciséis. En el Capítulo 5 se presentan los visemas que se usan en este trabajo para representar los fonemas en español.

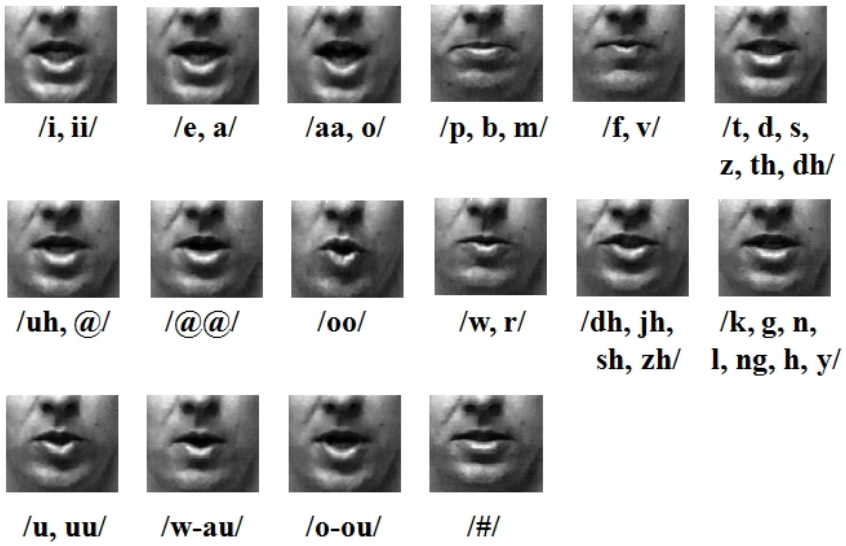

Figura 3-11: Correspondencia fonema-visema en inglés propuesto por Tony Ezzat y Tomaso Poggio.

Aunque el conjunto de visemas que presentan contiene fonemas que no pertenecen al español, se puede ver que hay consonantes que caen perfectamente en la clasificación de consonantes de acuerdo a su punto de articulación que se discute al inicio de este capítulo. Por ejemplo, las labiales /b/, /p/ y /m/ caen en la misma categoría de visema, igual que las labiodentales /f/ y /v/. Debe notarse que Ezzat y Poggio colocan en la misma clasificación de visema a las velares /k/ y /g/ con las alveolares /n/ y /l/, que de acuerdo a la clasificación de sonidos expuesta

aquí no sería posible, ya que tienen distintos puntos de articulación. Una vez más hay que hacer notar que la proyección entre fonemas y visemas depende del idioma en el que se haga y también de la finalidad de la animación, ya que si sólo es para animar un personaje de caricatura el conjunto de visemas no debe ser tan exacto a comparación de una animación que pretende comunicar con exactitud algún mensaje a través de los labios.

El presente capítulo establece las bases del proceso de sincronización del movimiento de labios en el habla del español. La correspondencia entre fonemas y visemas es la clave para obtener, a partir de un texto en español, los fotogramas de la animación. El proceso de silabificación, como se verá en el Capítulo 5, es esencial para asignar los tiempos a las transiciones entre visemas y para lograr un efecto de sonoridad creciente alrededor de los núcleos silábicos. En el siguiente capítulo se discuten, de manera formal, los objetivos generales y específicos de la tesis, así como los requerimientos funcionales y no funcionales del sistema a desarrollar.

Capítulo 4

Planteamiento del Problema

El Capítulo 3 concluye el marco teórico del proyecto, ahora procederemos a establecer los objetivos generales y específicos del trabajo. En este capítulo se discute además, el origen del problema y la motivación detrás de la investigación. Se presentan también, algunos trabajos realizados previamente por otros investigadores en el estado de Puebla, como argumentación de la tesis. Finalmente, se exponen los requerimientos funcionales y no funcionales del sistema a desarrollar.

4.1. Objetivos Generales

En este trabajo se pretende desarrollar un sistema que permita animar un rostro tridimensional a partir de un conjunto inicial de modelos 3D, que represente distintas expresiones faciales y distintos visemas (el término *visema* se discute en la sección 3.4). El sistema recibirá un texto en español y generará la animación del rostro de tal manera que se logre la apariencia de que está pronunciando las palabras del texto. Para ello, el texto pasará por un proceso de silabeo, ya que fonológicamente los segmentos (consonantes y vocales) se agrupan en sílabas. Una vez que se tienen las sílabas, éstas serán descompuestas en fonemas, que serán proyectados a su conjunto de visemas correspondiente para producir la animación.

4.2. Objetivos Específicos

- Crear la animación de expresiones faciales y del movimiento de los labios, aplicando técnicas de metamorfosis sobre la geometría de un conjunto de modelos 3D de un rostro.
- Implementar un sistema de silabeo y de análisis fonológico que satisfaga las necesidades del proyecto.
- Establecer una correspondencia entre fonemas y visemas propia del idioma español.
- Lograr que el sistema genere una animación suave e ininterrumpida, evaluando el desempeño del programa y su tiempo de ejecución para determinar si es posible realizar la animación en tiempo real, o si es necesario guardar los fotogramas para posteriormente procesarlos y crear una secuencia de video que pueda ser reproducida.
- Sincronizar la animación con el audio correspondiente a la pronunciación del texto de entrada.

4.3. Origen del Problema

En diciembre de 2004 José Leopoldo Díaz Alonso obtuvo su título de Licenciado en Ciencias de la Computación por parte de la Benemérita Universidad Autónoma de Puebla al presentar la tesis titulada "Prototipo de un Silabario Multimedia en el Idioma Español-Mexicano", en ella expone la necesidad de aplicar la computación en el campo de la educación especial, en particular la de personas con hipoacusia en educación preescolar y educación primaria [52].

La hipoacusia es la disminución del nivel de audición de una persona por debajo de lo normal y es tan amplio el campo de la pérdida de la audición que para facilitar su comprensión se puede clasificar en tres tipos, por el momento de adquirirla, por la localización de la lesión y por el grado de la pérdida auditiva. El momento en que se obtiene tiene importantes consecuencias en la adquisición del lenguaje oral, debido a que un niño puede adquirir sordera antes de hablar (prelocutiva) o después (poslocutiva), lo que determina gran parte de su tratamiento y rehabilitación. En ocasiones las personas que han perdido la audición siendo adultas o después de los 18 años, son llamadas sordos postvocacionales, las cuales pueden tener serios problemas

tanto personales como profesionales. En circunstancias más difíciles, se encuentran aquellas que han perdido la audición en edades tempranas, ya que no solamente van a tener problemas con el aprendizaje del lenguaje, sino también debido al menor número de oportunidades para interaccionar con personas oyentes en ambientes prevocacionales, estará en mayores desventajas en el área de las relaciones personales para enfrentarse al campo de trabajo [53].

Además de escuchar, la lectura labiofacial es otra manera para que el niño aprenda a entender el lenguaje hablado, al observar la boca, las expresiones faciales y los gestos de la persona que habla. El proyecto de Díaz Alonso consistió en mostrar videos de la boca humana, enfocándose al movimiento de los labios correspondiente a la sílaba indicada en el programa de aplicación, el cual funciona interactivamente tomando como dato de entrada el texto a silabear. Díaz Alonso propone como trabajo a futuro que el usuario pueda escoger entre varios modelos de la persona que habla (por ejemplo: hombre-mujer, niño-niña, etc.), lo que implicaría tener varias personas grabadas en video con cada una de las más de 2300 sílabas.

En nuestro proyecto se pretende desarrollar un programa parecido al del trabajo de Diaz Alonso, pero sin trabajar con personas reales grabadas en video sino con modelos de rostros tridimensionales que serán animados para lograr la apariencia de que están hablando. Con ello se logra no sólo evitar la necesidad de tener una base de datos relativamente grande de videos sino también, se podría diseñar un sistema con varios tipos de modelos de rostro, aunque en un principio se trabajará solamente con uno.

4.4. Motivación

Al madurar y hacerse más accesible la tecnología de software y hardware para la animación facial, se pueden explorar diversas aplicaciones de creciente complejidad. Mientras que la industria del entretenimiento y la visualización científica-médica seguirán generando aplicaciones cada vez más sofisticadas, se esperan avances significativos en el amplio campo de la interacción humano-computadora. Por ejemplo, campos emergentes y activos se relacionan con la computación-afectiva (*affective computing*) y agentes personificados de conversación (*embodied conversational agents*) que entre otros tratan el problema de "intentar mejorar las interacciones entre los humanos y la tecnología mediante el desarrollo de sistemas artificiales que respondan

al humor y emociones del usuario humano" [54]. Estas interacciones suceden cada vez más entre humanos y agentes humanoides animados de conversación (*humanoid animated conversational agents*). Ha sido mostrado concluyentemente que tal interacción es mejorada por agentes animados que exhiben habilidades sociales-emotivas y/o asociaciones socio-culturales y parecen tener vida (*lifelike*) [32]. Para lograr estas metas el modelado y animación del rostro, expresiones faciales, voz, estilo visual y la personalidad resultante de tal agente juegan un papel vital y ofrecen una amplia oportunidad para una investigación multi-disciplinaria.

Es por eso que se considera importante el unificar los conceptos en computación aplicados en el campo de la educación especial, para comprender el alcance y limitaciones que se tienen con respecto a estos campos. En particular al de personas hipoacúsicas en educación preescolar y educación primaria. Sabemos que la audición como vía habitual para adquirir el lenguaje y desarrollar el habla, es el atributo principal que permite la adquisición de todo proceso psicológico, en particular la comunicación. Actualmente las estadísticas denotan que la adquisición de la hipoacusía es factorial, lo que hace pensar en la planificación de estrategias diversas y operacionales para la rehabilitación o en el mejor de los casos de la reeducación auditiva, esto es, crear los programas específicos con estímulos adecuados para su futura reincorporación a su medio ambiente social-laboral [52].

El concepto que se utilizó en el XII Censo General de Población y Vivienda 2000 del INEGI[1] [53], considera a las personas con discapacidad auditiva como aquellas que presentan pérdida o restricción de la capacidad para recibir mensajes verbales u otros mensajes audibles. La Organización Mundial de la Salud (OMS), señala que entre uno y dos por cada mil de los recién nacidos llegan al mundo siendo sordos profundos o severos. Para el año 2001 se estimó que en el mundo había 250 millones de personas con discapacidad auditiva.

Entre los esfuerzos que se han realizado en favor de las personas con discapacidad, se encuentran el Programa de Acción Mundial para las Personas con Discapacidad, cuyo objetivo es promover medidas eficaces para la prevención de la incapacidad, la rehabilitación y la integración de los impedidos en la vida social y en el desarrollo. Las Normas Uniformes sobre la Igualdad de Oportunidades para las Personas con Discapacidad, cuya finalidad es buscar que los estados del mundo garanticen que niñas y niños, mujeres y hombres con discapacidad,

[1] Insituto Nacional de Estadística y Geografía en México.

en su calidad de miembros de sus respectivas sociedades, puedan tener los mismos derechos y obligaciones que el resto de la población. En la región, la Convención Interamericana contra la Discriminación de las Personas con Discapacidad, tiene como objetivo la prevención y eliminación de todas las formas de discriminación contra las personas con discapacidad y propiciar su plena integración en la sociedad mediante la adopción de medidas de carácter legislativo, social, educativo, laboral o de cualquier otra índole, necesarias para eliminar la discriminación contra las personas con discapacidad y propiciar su plena integración en la sociedad. La Organización Internacional del Trabajo (OIT) a través del Convenio 159 sobre Readaptación Profesional y el Empleo de Personas Inválidas de 1983, estableció un acuerdo internacional que define la política destinada a asegurar que existan medidas adecuadas de readaptación profesional al alcance de todas las categorías de personas inválidas (discapacidad) y a promover oportunidades de empleo en el mercado regular, así como establecer un catálogo de derechos que deben ser gozados por cualquier trabajador con discapacidad en cualquier parte del mundo [53].

México ha ratificado algunos de estos esfuerzos, lo que trajo como consecuencia que a mediados de los años noventa se iniciara en el país un esfuerzo más coordinado entre distintos sectores y se incluyera en la política pública el tema de la discapacidad.

Los resultados del XII Censo General de Población y Vivienda 2000 del INEGI (Figura 4-1), reportaron casi tres personas con discapacidad auditiva por cada mil habitantes en el país, esto significa alrededor de 281 mil personas, de las cuales 31.2 % residían en el medio rural. Se observa al interior de las entidades federativas que la de mayor prevalencia de discapacidad auditiva fue Yucatán, al contar con 4.4 personas con discapacidad auditiva por cada mil habitantes, seguida por Zacatecas e Hidalgo ambas con cuatro personas. En contraste, Baja California (1.7), Chiapas (1.9) y Quintana Roo (2), fueron las entidades de menor prevalencia con esta discapacidad.

Como se ha mencionado anteriormente, el momento en que se adquiere la discapacidad, tiene importantes consecuencias en el aprendizaje del lenguaje oral, debido a que un niño puede adquirir sordera antes de hablar, tener sordera prelocutiva, o bien, después de hablar, tener sordera poslocutiva; dependiendo del tipo de sordera que adquiera se determina el tipo de tratamiento y rehabilitación. Existen diversas causas por las que se puede adquirir este tipo de discapacidad, en México 38.2 % de esta población declaró haber adquirido su discapacidad

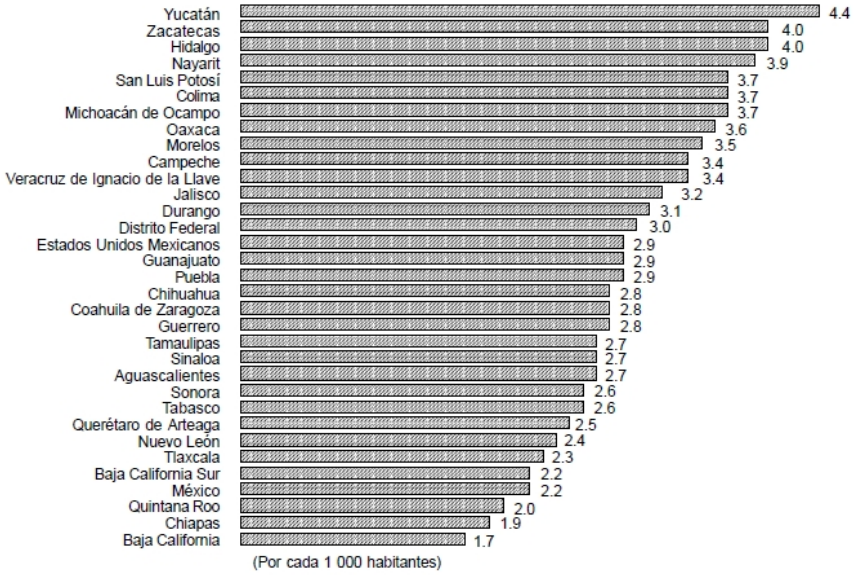

Figura 4-1: Prevalencia de discapacidad auditiva por entidad federativa, 2000. FUENTE: **INEGI**. *XII CGPV 2000. Base de datos.*

debido a la edad avanzada (también conocida como presbiacusia), ésta se relaciona con un proceso de deterioro físico o mental que acompaña al envejecimiento. La presbiacusia afectó más a las mujeres (40.6 %) que a los hombres (36.3 %). Las enfermedades resultaron ser la segunda causa de la discapacidad auditiva con 25.5 %, entre las que se pueden encontrar, las de la niñez que resultan en fiebres muy altas e infecciones del oído. Como en el caso anterior, la población femenina resultó más afectada (27.1 %) que la masculina (24.3 %). De las personas con esta discapacidad, 16.2 % de los casos tuvieron su origen al rededor del nacimiento, relacionados en gran medida con factores hereditarios, enfermedades eruptivas de la madre (rubéola, sarampión, varicela, etc.), ingestión de medicamentos durante el embarazo, incompatibilidad sanguínea, parto prematuro, uso de maniobras, fórceps mal aplicados; en los hombres esta situación representó 15.5 % y en el caso de las mujeres 17.1 por ciento. Finalmente, 11.8 % de las personas con discapacidad auditiva declararon como causa de ella algún accidente (contusiones, conmociones, fracturas); se presentó con mayor frecuencia en los hombres (15.3 %) que en las mujeres (7.5 %). Cabe señalar que para el año 2000, la primer causa de muerte en la población entre 1 y 29 años fue producto de accidentes [53].

Según la Ley General de Educación (1993), la educación es el medio fundamental para adquirir, transmitir y acrecentar la cultura; es el proceso permanente que contribuye al desarrollo del individuo y a la transformación de la sociedad, de igual forma, la educación especial para las personas con discapacidad debe ser impartida a la población de acuerdo a sus propias condiciones de manera adecuada y con equidad social.

Los derechos humanos de las personas sordas en términos de educación, se refieren al derecho a la oralización, manejo del lenguaje de señas y que sus estudios sean interpretados en lenguaje de señas por personal calificado. La oralización implica un diagnóstico bien realizado, adaptación de un auxiliar auditivo, terapia del lenguaje realizada y asistencia a una escuela regular. El lenguaje de señas es la lengua de las personas sordas, no es una lengua universal, cada país tiene su propia lengua, en el caso de México se conoce como Lengua de Señas Mexicana (LSM). Para lograr una comunicación efectiva se requiere de un intérprete, el cual constituye un puente entre las personas con discapacidad auditiva y las personas que no entienden este lenguaje. En este sentido, la finalidad de la educación especial consiste en lograr la autonomía personal y adaptación social de las personas con discapacidad. Desde tal perspectiva, son metas a lograr la integración escolar, la integración laboral y la integración social. Hoy en día se entiende por educación especial, el conjunto de apoyos y adaptaciones que ha de ofrecer la escuela para que el alumno integrado pueda seguir su proceso en el desarrollo y en el aprendizaje.

La asistencia escolar es un parámetro importante para medir la integración escolar, al poner en evidencia las oportunidades de educación con que cuenta la población con discapacidad auditiva; en este sentido, 43.2 % de las personas de 6 a 29 años con este tipo de discapacidad asistían a alguna escuela. Se observa que a medida que aumenta la edad disminuye la asistencia escolar: 76 de cada 100 personas de entre 6 y 9 años asistían a un centro educativo, 74 de entre 10 y 14 años, 36 de entre 15 y 19 años, 11 de entre 20 y 24 años y sólo 5 de los de 25 a 29 años de edad. Una proporción importante de las personas con discapacidad auditiva (55.2 %) no asistía a la escuela. El abandono escolar puede iniciar desde el primer grado que se cursa; para ello, es necesario haber asistido en algún momento a la escuela; cabe recordar que el ingreso a la educación primaria es a partir de los seis años, de modo que se considera el abandono escolar un año después, a los siete años. En este contexto, de las personas de 7 a 29 años que no asistieron a la escuela, 19.3 % no habían asistido nunca a un plantel educativo, por su parte,

de los que abandonaron la escuela (72.1 %), 26 % se debió por falta de dinero, 21.6 % porque no les gustó estudiar y el resto por diversas razones. Una condición esencial que se relaciona con el acceso educativo, es adquirir la habilidad de lecto-escritura; en este sentido, 70.6 % de las personas de 8 a 14 años con discapacidad auditiva contaban con estas habilidades, dato que resulta importante si se considera que 56.5 % de la población con discapacidad en general sabía leer y escribir.

A una persona que tiene 15 años o más sin la habilidad de leer y escribir se le considera analfabeta, en el 2000, 66.7 % de las personas con discapacidad eran analfabetas al momento del censo, contra 34.8 % de las personas con discapacidad auditiva; se observa una brecha significativa por sexo, 29.2 % de los varones eran analfabetas, mientras que la población femenina concentró 41.6 por ciento. El nivel de escolaridad alcanzado por estas personas, presenta diferencias por sexo que evidencian desventaja de las mujeres frente a los hombres: 35.7 % de los varones no tuvieron instrucción, 31 % no habían completado la primaria y sólo 14.7 % la terminó; por su parte en las mujeres 44 % no fueron instruidas, 25.9 % no completaron la primaria y apenas 13.9 % sí concluyó. Por otra parte, de las personas con discapacidad auditiva de 15 años y más, apenas 5.9 % había completado la educación básica, 4.4 % tenían educación media superior y 3.2 % lograron estudios superiores o de posgrado.

La información sobre las características educativas de las personas con discapacidad auditiva, muestra que cuentan con una asistencia escolar regular, un importante grado de deserción y un nivel de instrucción mediano, factores que determinan que el promedio de escolaridad sea 3.4 años por persona, 3.6 años en el caso de los hombres y 3.1 años en el caso de las mujeres. Cabe recordar que una parte importante de las personas con discapacidad auditiva, adquirieron su discapacidad en edad avanzada y pocos la obtuvieron al nacer; la estructura por edades muestra un panorama que permite concluir que gran parte de estas personas, adquirieron la discapacidad a una edad mayor [53].

Existe una gran variedad de métodos y técnicas que se usan en la educación especial. Muchos profesores usan una combinación de métodos. Existen dos objetivos al trabajar con la lectura de labios, por un lado, que el alumno capte información a través de los labios de los interlocutores que sólo manejan lenguaje oral con lo que se potencia la posibilidad de comunicarse con la mayoría de los oyentes, por el otro manejar correctamente la estructura de la lengua oral lo que

le resulta ser un elemento facilitador en el momento de realizar actividades de lecto-escritura.

4.5. Trabajos Relacionados

En el Estado de Puebla se han llevado a cabo varias investigaciones relacionadas con la educación especial y en particular, en relación a la educación y rehabilitación de niños con capacidades auditivas diferentes. En esta sección se presenta una breve descripción de algunos de estos trabajos, con el objetivo de poner en contexto el presente proyecto.

En el año 2000 la Ingeniera en Ciencas de la Computación Leticia Ruiz Flores presentó un sistema de apoyo para terapia de lenguaje como su tesis de licenciatura "Icatiani: Interfaz para Apoyo en Terapia de Lenguaje" [55] bajo la dirección de la Dra. Ingrid Kirschning en la Universidad de las Américas Puebla. Su sistema busca brindar apoyo a maestros y terapeutas en la creación de lecciones basadas en discriminación de dos fonemas o palabras, dependiendo de lo que se quiera practicar. A la interfaz se incorpora una cara virtual llamada Baldi, la cual fue desarrollada en la Universidad de California en Santa Cruz. Baldi mueve los labios en sincronía con la producción de sonido de un sintetizador o a la par de un archivo de sonido, con el objetivo de que el niño practique la lectura labio-facial y el punto de articulación de los diferentes sonidos. Desgraciadamente, como indica Ruiz Flores, Baldi funciona muy bien en inglés, pero se presentan ciertos errores en algunos fonemas en español, esto se debe a que fue desarrollado originalmente para el idioma inglés. La Figura 4-2 muestra el modelo 3D de Baldi con distintas opciones de presentación.

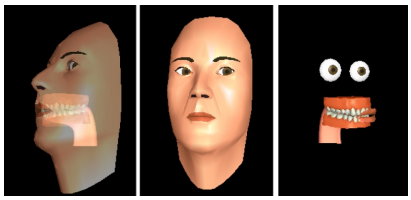

Figura 4-2: Sistema de animación facial, Baldi, desarrollado por Perceptual Science Lab, University of California, Santa Cruz.

Otro trabajo relevante para la educación especial en el estado, es el de Mónica Limón Rosas, presentado como su tesis de licenciatura en Ciencias de la Computación en la Benemérita

Universidad Autónoma de Puebla bajo la coordinación del M.C. José Esteban Torres León, titulada "Intranet para Personas con Sordera"[56]. Su sistema proporciona un conjunto de herramientas que realizan la traducción de lenguaje de señas a lenguaje escrito. También, al seleccionar una seña determinada, el sistema proporciona su lenguaje de señas, lectura labio-facial, texto relacionado e imagen asociada. El sistema se compone de un diccionario, que provee el lenguaje escrito, la imagen asociada y lenguaje de señas (en forma de animación) de una seña en particular; un traductor, para ayudar a la organización lógica del lenguaje y un *chat* que favorece el desarrollo afectivo social del niño. Su sistema fue probado con niños con sordera teniendo un éxito razonable. Las principales desventajas que menciona Limón Rosas es el vocabulario limitado con el que cuenta el sistema, pero éste cuenta con opciones para incrementarlo. La Figura 4-3 muestra la interfaz de su programa.

Figura 4-3: Traductor de lenguaje de señas desarrollado en la Universidad Autónoma de Puebla (Limón Rosas, 2003).

Finalmente, como se expone en la sección 4.3, el trabajo que dio origen a esta tesis es el de Leopoldo Díaz Alonso. Su sistema consiste en un silabario multimedia que permite al usuario introducir un texto en español y observar una secuencia de video que corresponde a la pronunciación del texto de entrada. Para lograr esto, Díaz Alonso grabó videos de él mismo pronunciando distintas sílabas por separado. Al procesar el texto de entrada, el programa muestra los videos correspondientes en secuencia para representar la pronunciación del texto. Uno de los puntos interesantes del trabajo es su sencilla interfaz, además de que permite a los niños jugar a identificar por medio de la lectura de labios la palabra que se está pronunciando. Al niño se le muestra una secuencia de videos que corresponden a un texto escrito por una

tercera persona, el cual el niño desconoce, él deberá poner atención en el movimiento de labios y tratar de identificar la palabra. El sistema es sencillo y útil, principalmente gracias a que Díaz Alonso trabajó de cerca con los usuarios finales desde el principio, pudiendo así satisfacer sus necesidades.

La principal desventaja del trabajo de Díaz Alonso es el hecho de que se deben tener una gran cantidad de videos para poder abarcar todas las posiblidades de sílabas del español, además de que los usuarios no tienen ningún tipo de interacción con las secuencias de video ya que han sido grabadas previamente. Para solucionar esto, en este proyecto se propone usar modelos 3D para generar la animación de la pronunciación del texto, evitando así la necesidad de videos pregrabados y permitiendo al usuario una mayor interacción con la animación, al añadir expresiones faciales y otras opciones que ofrecerá sistema.

4.6. Especificaciones Funcionales

Como se expuso previamente en los objetivos generales del trabajo, se pretende desarrollar un sistema que permita animar un rostro tridimensional a partir de un texto en español, de tal manera que se logre la apariencia de que está pronunciando las palabras del texto. En esta sección se especifican los requerimientos funcionales y no funcionales del sistema.

4.6.1. Requerimientos Funcionales.

Los requerimientos funcionales son aquellos que describen los servicios que debe proveer el sistema, cómo debe comportarse en situaciones particulares y en relación a datos de entrada específicos, así como también algunas de sus restricciones [57]. Para el sistema que se va a desarrollar se establecen los siguientes requerimientos funcionales:

- El sistema recibirá como entrada un texto en español. Éste podrá contener acentos y signos de puntuación, aunque éstos últimos serán ignorados al generar la animación.
- El sistema aceptará ciertas secuencias de caracteres para modificar la expresión del rostro, por ejemplo =) para sonreír, aunque la interfaz contará con una serie de botones parecidos a los de los servicios de mensajería instantánea en línea para incluir iconos (comúnmente llamados *emoticons*), que representen de manera visual a las distintas expresiones faciales

(en el Capítulo 5 se discuten a detalle las secuencias de caracteres para las expresiones así como sus iconos respectivos).

- El usuario podrá controlar el tiempo asignado por sílaba, especificándolo en milisegundos.
- En cuanto a la sincronización con audio, el usuario contará con una herramienta para ingresar datos que permitan al sistema sincronizar la animación con un archivo de audio existente (la naturaleza de estos datos será explicada en el Capítulo 5).
- La salida del sistema será la animación del rostro tridimensional correspondiente a la pronunciación del texto de entrada. La animación será generada no sólo a base del texto de entrada, sino también de los tiempos asignados y la secuencia de expresiones que se indiquen para el rostro. Idealmente, la animación se mostrará después de un corto tiempo de procesamiento, es decir, el usuario podrá ver la animación casi enseguida de haber presionado el botón para iniciar el proceso. Sin embargo, se contempla la posibilidad de que se genere un video para poder reproducirlo después, agregando en la interfaz un control para que el usuario decida si se guarda el video o no.
- La interfaz del sistema proveerá algunas funciones para modficar la apariencia de la visualización 3D: cambiar el color de piel del rostro, el color de ojos, el color de fondo y la textura del fondo. Todo esto con el objetivo de que exista mayor interacción entre el usuario y la animación. Uno de las funciones que no se incluirá en esta versión del sistema es la de cambiar de modelo de rostro, ya que se requiere otro conjunto de modelos 3D y no se cuenta con él.
- El sistema de visualización 3D permitirá rotar y acercar la cámara al modelo, para verlo de distintos ángulos.

4.6.2. Requerimientos no Funcionales

Los requerimientos no funcionales, como su nombre sugiere, son requerimientos que no se ocupan directamente con las funciones específicas que provee el sistema. Pueden relacionarse con propiedades del sistema como la seguridad, tiempo de respuesta y espacio de almacenamiento. Alternativamente, pueden definir restricciones en los servicios o funciones que ofrece el sistema.

Por ejemplo, restricciones de tiempo o del proceso de desarrollo [57]. Se definen los siguientes requerimientos no funcionales:

- El sistema trabajará con un conjunto estático de modelos 3D para el proceso de metamorfosis que serán cargados a memoria al momento de iniciar el programa.
- El analizador fonológico deberá silabificar adecuadamente las palabras del español para lograr una animación más realista de su pronunciación.
- Para lograr la animación en un corto tiempo de procesamiento, el programa deberá realizar las operaciones de combinación de modelos, interpolación y visualización de manera eficaz para que la animación no se vea entorpecida.
- Para el caso de la generación del video, se deberá contar con un espacio de memoria relativamente amplio para poder guardar cada uno de los fotogramas de la secuencia, el espacio estará relacionado directamente con la duración de la animación.
- La interfaz deberá ser sencilla de utilizar, pues de manera ideal se pretende que lo puedan usar niños de primaria.
- El sistema debe ser portable, sobre todo no debe depender de una tarjeta de gráficos en particular.

Una vez establecidos los objetivos y requerimientos del sistema, podemos seguir con el proceso de diseño e implementación que se discute en el siguiente capítulo.

Capítulo 5

Diseño e Implementación

La técnica de animación facial que se utiliza en este trabajo es la de fotogramas claves con interpolación lineal. Como se explicó en el Capítulo 2, esta técnica consiste en definir dos posturas clave del rostro (en este caso, dos modelos tridimensionales) e interpolar entre las posiciones de los vértices en cada una de ellas. Por lo tanto, el primer problema a resolver es la creación de un conjunto de modelos 3D que representen distintas expresiones faciales (incluyendo visemas) y que sean adecuados para realizar la metamorfosis entre ellos. Una vez que se tiene este conjunto de modelos 3D, es necesario definir un método que nos permita generar nuevas expresiones faciales a partir de las originales, para así poder hacer combinaciones de los modelos del conjunto inicial. Para ello, se utiliza la metamorfosis ponderada, donde un modelo 3D se puede generar al combinar otros modelos que tienen asignados diferentes pesos. Así, cada vértice del nuevo modelo tiene una posición igual a la combinación lineal de su posición correspondiente en cada uno de los modelos que tienen un peso distinto de cero. Con lo anterior, es posible generar la animación del rostro utilizando distintas expresiones faciales (y distintos visemas en el caso del movimiento de labios).

Ahora bien, el sistema debe tomar como entrada un texto en español y a partir de él, generar la animación del rostro. Para lograr esto, se necesita diseñar un analizador fonológico que, a partir del texto de entrada, identifique los elementos esenciales para la animación, a saber, la secuencia de visemas correspondiente y sus tiempos respectivos. También, para añadir expresividad al rostro durante la animación, se debe diseñar e implementar un sistema de interacción tipo *script*, con el cual el usuario pueda modificar la expresión o el sentimiento con

el cual se pronuncia una parte determinada del texto.

Se pretende que la animación se muestre al usuario al mismo tiempo que se realiza la interpolación entre visemas y sin necesidad de grabar la secuencia en video. Sin embargo, como se verá más adelante, grabar en video la animación para poder reproducirla sin necesidad de volver a procesar los datos de entrada tiene algunas ventajas: el video muestra una animación más fluida y además se puede reproducir las veces que se desee.

Por lo tanto, el proceso de diseño e implementación del sistema puede dividirse en las siguientes etapas:

- Creación de los modelos 3D para la metamorfosis.
- Aplicación de interpolación y metamorfosis ponderada a los modelos 3D originales para generar la animación.
- Diseño e implementación de un analizador fonológico para procesar un texto en español y obtener los visemas correspondientes y sus tiempos para la animación.
- Diseñar un modo de interacción tipo *script* para modificar la expresión del rostro durante la animación.
- Presentar de la mejor manera posible la animación al usuario.

En primer lugar, es necesario establecer las herramientas de trabajo (lenguaje de programación, librerías, etc.) que se utilizarán. Posteriormente, se discute el diseño e implementación de cada una de las etapas anteriores, finalizando con un esquema del sistema final.

5.1. Ambiente de Desarrollo

Al decidir qué lenguaje de programación utilizar, uno de los factores más importantes que influyó en la decisión, fue la cantidad de documentación disponible con respecto al manejo de gráficos 3D. Existen varios libros y bastante información en la red sobre la interfaz de programación Java 3D (**véase Apéndice B**), por lo que se determinó trabajar con Java como lenguaje de programación. Aunque Java tiene muchas ventajas, como su paradigma orientado a objetos, soporte entre plataformas, reutilización de código, etc. existen algunas críticas con

respecto a su desempeño en áreas como la programación de videojuegos o los gráficos 3D. Algunos sostienen que Java es demasiado lento, en particular se refieren a los componentes de interfaz gráfica de usuario Swing, que son creados y controlados por Java, con poco soporte del sistema operativo; esto incrementa su portabilidad y hace que sean más controlables dentro de un programa en Java. Sin embargo, la velocidad de ejecución se ve comprometida debido a que Java impone una capa extra de procesamiento sobre el sistema operativo. Ésta es una de las razones por la cual muchas aplicaciones de juegos todavía utilizan el *Abstract Windowing Toolkit* (AWT) original, aunque gran parte del procesamiento de gráficos es manejado por hardware o software externo a Java. Por ejemplo, Java 3D delega sus tareas de *rendering* a OpenGL o DirectX, que puede emular capacidades de hardware. Otra de las críticas a Java es su nivel de abstracción que afecta el desempeño de gráficos de alta velocidad, pero desde la versión J2SE 1.4 se introdujo el modo FSEM (*Full-Screen Exclusive Mode*), que suspende el ambiente de ventanas normal y le permite a una aplicación acceder al hardware de gráficos de manera más directa.

Java 3D puede cargar modelos externos a través de su interfaz **Loader** y la clase **Scene**. El paquete de utilidades de Java 3D incluye dos subclases de **Loader** diseñadas para formatos específicos de archivo: **Lw3dLoader** maneja archivos de escenas 3D de *Lightwave* y **ObjectFile** procesa archivos OBJ de *Wavefront*. Una tercera subclase, **LoaderBase**, implementa la interfaz **Loader** de forma genérica para alentar el desarrollo de clases para otros formatos 3D.

Existe una gran variedad de cargadores de Java 3D escritos por terceros para diferentes formatos de archivo. En este trabajo se utiliza el paquete **NCSA Portfolio**, que soporta varios formatos incluyendo 3D Studio Max (archivos 3DS), AutoCAD (DXF), Digital Elevation Maps (DEMs), TrueSpace (COB) y VRML 97 (WRL). Las desventajas de este paquete son su avanzada edad y su soporte relativamente simple de los formatos: frecuentemente sólo se puede cargar la geometría y los colores de la figura, sin texturas, comportamientos, o luces [58].

5.2. Creación y Manejo de los Modelos 3D

Modelar un rostro tridimensional no es una tarea sencilla, incluso usando programas de diseño asistido por computadora se requiere de gran habilidad para lograr que tenga una buena

apariencia. Además, cuando el objetivo es la animación facial, se debe cuidar que el modelo tridimensional del rostro tenga una topología apropiada. Como ya se mencionó, la técnica de animación facial que se utilizará es la de interpolación de fotogramas clave, es decir, que a partir de un conjunto inicial de modelos 3D se realizará la metamorfosis entre ellos. Para esto, es necesario que los modelos 3D compartan información geométrica y de conectividad para que no haya problema al interpolar las posiciones de los vértices.

Inicialmente, se pensó en construir los modelos del rostro a partir de información visual de una persona real, sin embargo, el proceso hubiera tomado mucho tiempo y la calidad de los modelos probablemente no hubiera sido muy buena. La otra opción era modelar el rostro usando un programa comercial de modelado 3D (como Maya o 3dStudioMax), pero esto requiere de una gran habilidad, así que se decidió adquirir en la red[1] un conjunto de modelos 3D de un rostro. Este conjunto contiene: una cabeza con la expresión neutral (es decir, sin ninguna expresión), 22 modelos para las expresiones faciales y 16 modelos para los visemas. En la Figura 5-1 se muestran algunos de estos modelos, no se incluyen los 39 porque algunos sólo varían en la posición de los ojos, las pestañas o las cejas y las diferencias no son tan perceptibles. Los visemas que se muestran corresponden a fonemas del idioma inglés, así por ejemplo, el visema marcado como **I** en realidad correspondería a la pronunciación *ai* en español.

Los modelos se cargan en tiempo de ejecución utilizando la clase **ModelLoader** del paquete **NCSA Portfolio** y son guardados en una matriz de objetos de la clase **GeometryArray**. Esta clase pertenece a Java 3D, es una clase abstracta a partir de la cual se derivan varias clases para especificar un conjunto de primitivas geométricas. Un objeto **GeometryArray** contiene arreglos separados de los siguientes componentes de los vértices: coordenadas, colores, normales, coordenadas de textura y una máscara de bits que indica cuáles de estos componentes están presentes. Se utiliza una matriz de estos objetos debido a que los modelos 3D con los que se va a trabajar no están hechos de una sola pieza: la piel, los ojos, los dientes y la lengua son objetos independientes y cada uno posee su propia información geométrica y de conectividad.

A continuación se explicará el proceso de metamorfosis entre modelos y posteriormente se discutirá la proyección entre visemas y fonemas del español.

[1] Existen varias páginas en la red dedicadas a la compra y venta de modelos 3D creados por diferentes artistas, los modelos que se usan en este trabajo fueron adquiridos en www.the3dstudio.com y creados por *AnarchyRising*.

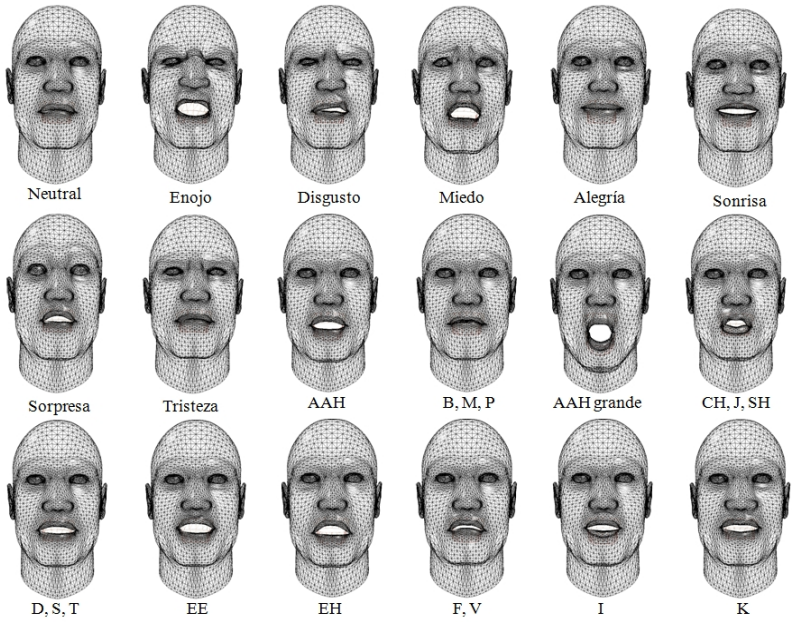

Figura 5-1: Conjunto original de modelos 3D. Solamente se muestran las principales expresiones faciales y un subconjunto de visemas, estos últimos corresponden al idioma inglés.

5.3. El Proceso de Metamorfosis

Para realizar la transformación entre una expresión facial y otra se utiliza la técnica de interpolación lineal simple, ésta consiste en generar los puntos intermedios de una línea recta a partir de los dos puntos que la determinan. En la metamorfosis 3D estos puntos existen en el espacio tridimensional y al formar parte de una superficie se logra el efecto de que ésta se deforma.

En la sección 2.3.1 se discute la técnica de interpolación lineal de manera general, veamos ahora cómo se aplica en la metamorfosis de modelos 3D. Un modelo 3D está compuesto por vértices y aristas (aunque existen otras formas de representación como se expuso en el Capítulo 1). Para transformar un modelo 3D en otro, se interpola sobre la posición de los vértices de los modelos, es decir, teniendo el k-ésimo vértice del modelo original Vo_k y su vértice correspondiente en el modelo destino Vd_k, se realiza una interpolación lineal entre las posiciones de estos vértices en el espacio tridimensional (ecuación 5.1), con $0 \leq k \leq m-1$, donde m es el número de vértices de los modelos 3D. Cabe señalar que para que la interpolación pueda realizarse, los

modelos tienen que tener el mismo número de vértices. El parámetro t de la interpolación es el tiempo, de esta forma, se logra la transición del modelo original al modelo destino a través del tiempo, con $0 \leq t \leq 1{,}0$.

$$V_k(t) = Vo_k + t \cdot (Vd_k - Vo_k) \tag{5.1}$$

Cuando la transición comienza, $t = 0{,}0$, por lo que la posición del vértice es la misma que en el modelo original, es decir, $V_k(0{,}0) = Vo_k$. Al finalizar la transición el valor de t es uno, $t = 1{,}0$ y la posición del vértice es la misma que en el modelo destino, $V_k(1{,}0) = Vo_k + 1{,}0 \cdot (Vd_k - Vo_k) = Vd_k + (Vo_k - Vo_k) = Vd_k$. Para cualquier valor de t entre 0.0 y 1.0, la posición del vértice se encuentra dentro de la línea recta formada por el vértice del modelo original y el vértice del modelo destino. Se debe enfatizar que, el parámetro t debe tomar valores entre cero y uno, por lo tanto, al establecer el tiempo de la animación se debe determinar un incremento de tiempo proporcional para que no salga de este rango. En la implementación, este incremento se establece como $dt = 1{,}0/nframes$, donde $nframes$ es el número de fotogramas que debe tardar la transición, así por ejemplo, si la animación debe tardar un segundo y se está trabajando a 24fps (fotogramas por segundo, *frames per second*), entonces el incremento de tiempo será $dt = 1/24$, si la transición ocurre en dos segundos entonces $dt = 1/48$.

Java 3D contiene algunas clases que ayudan a realizar animaciones, una de ellas es la clase abstracta **Behavior**. Esta clase provee los medios para animar objetos, procesar la entrada del teclado y el ratón, reaccionar a movimientos, asó como habilitar y procesar eventos de selección. Los dos métodos principales de esta clase son **initialize()** y **processStimulus()**, el primero se utiliza para inicializar las variables de estado del objeto **Behavior** y establecer las condiciones que deben cumplirse para que se ejecute el método **processStimulus()**, conocidas como **WakeupConditions**. Existen varias condiciones que se pueden establecer, en este caso se usó la condición **WakeupOnElapsedFrames**, que especifica que Java 3D debe despertar al objeto **Behavior** después de que haya procesado el número especificado de fotogramas. La clase que se implementó, llamada **MorphBehavior**, realiza la interpolación dentro del método **processStimulus()**, así se va deformando el modelo 3D con cada fotograma que es procesado, permitiendo que el tiempo de animación se mida en fotogramas por segundo (fps).

Otra clase que ofrece Java 3D para animar objetos 3D es la clase **Morph**. Un objeto de

esta clase crea la geometría de un objeto visual interpolando sobre un conjunto de objetos de la clase **GeometryArray**. La combinación de los objetos se hace de manera ponderada, con la restricción de que la suma de los pesos de cada uno de los objetos debe ser igual a uno. Esto presentó un problema, ya que con esta clase no se logra combinar totalmente los modelos, por ejemplo, teniendo un modelo 3D del rostro con el ojo izquierdo cerrado y otro con el ojo derecho cerrado se esperaría que su combinación fuera un modelo con los dos ojos cerrados, sin embargo, lo que se obtiene con la clase **Morph** es un modelo con los dos ojos medio cerrados, ya que el peso de cada modelo debe ser igual a 0.5 para que su suma sea igual a 1.

Para solucionar este problema se diseñó una clase con el nombre de **MyMorph** que utiliza también la metamorfosis ponderada pero tomando un modelo 3D como base para la combinación. En este caso, el modelo base es el del rostro sin expresiones, así, se toman los modelos que se vayan a combinar y se calcula la diferencia entre cada uno de los vértices de estos modelos y los del modelo base, obteniendo un desplazamiento para cada punto en el espacio 3D que posteriormente se suma a las posiciones de los vértices del modelo base, obteniendo una nueva geometría que corresponde a la combinación lineal de los modelos elegidos. La Figura 5-2 muestra este proceso: se empieza con un modelo base (Figura 5-2a), se seleccionan dos modelos (Figura 5-2b y 5-2c) para combinarse (pueden ser más de dos modelos), se obtienen las diferencias entre los vértices de estos modelos y los del modelo base (Figura 5-2d y 5-2e) y se suman las diferencias al modelo base (Figura 5-2f). En este ejemplo, el modelo base es el rostro sin expresiones y se pretende combinar el modelo del rostro con el ojo izquierdo cerrado con el modelo que tiene el ojo derecho cerrado. Como se puede observar, estos modelos difieren del modelo base únicamente en los vértices de los ojos (el izquierdo y el derecho respectivamente), mientras que los vértices del restro de la geometría se encuentran en la misma posición que en el modelo base. Así, al calcular las diferencias solamente se obtendrán desplazamientos diferentes de cero para los vértices de los ojos. Note que para el modelo con el ojo izquierdo cerrado, los vértices del ojo derecho están en la misma posición que en el modelo base, por lo que su desplazamiento será cero. Lo mismo sucede para el modelo con el ojo derecho cerrado, así, al sumar estas diferencias al modelo base, los desplazamientos de un ojo se suman sin afectar los desplazamientos del otro y se genera el modelo con los dos ojos cerrados.

El programa va guardando los desplazamientos para cada vértice en un arreglo de puntos

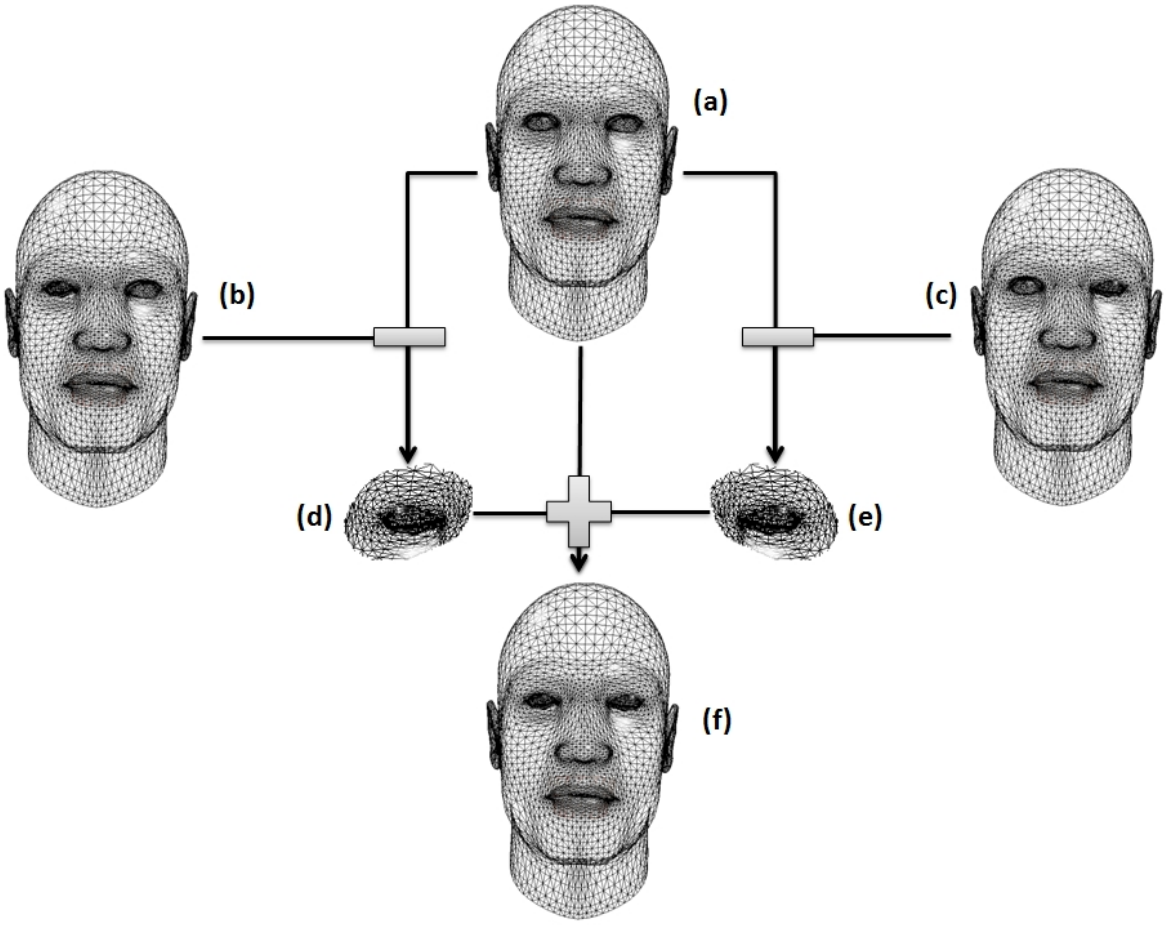

Figura 5-2: Combinación de modelos 3D. (a) Modelo base, (b) rostro con el ojo derecho cerrado, (c) rostro con el ojo izquierdo cerrado, (d) diferencia entre el modelo base y el modelo (b), (e) diferencia entre el modelo base y el modelo (c), (f) modelo resultante de sumar las diferencias al modelo base.

3D, el desplazamiento total para un vértice es la suma ponderada del desplazamiento obtenido entre el modelo base y cada uno de los modelos que se van a combinar, así, se pueden generar diferentes combinaciones entre los modelos, obteniendo una gran variedad de expresiones. La ecuación 5.2 muestra la forma de obtener el k-ésimo vértice V del nuevo modelo. Primero, se obtiene la diferencia de la posición del k-ésimo vértice en cada uno de los n modelos originales (Vi_k, con $0 \leq i < n$) y su posición en el modelo base (Vb_k), multiplicando esta diferencia por el peso del i-ésimo modelo (ω_i). La suma ponderada de estas diferencias se suma a la posición del vértice en el modelo base (Vb_k), esto se hace para cada uno de los vértices del modelo, que en nuestro caso son 36,840, es decir, $0 \leq k \leq 36,839$ tomando como índice inicial el cero.

$$V_k = \sum_{i=0}^{n}((Vi_k - Vb_k) \cdot \omega_i) + Vb_k \tag{5.2}$$

5.4. Proyección entre Fonemas y Visemas

Para generar la animación del movimiento de los labios, es necesario establecer una correspondencia entre los fonemas del español y un conjunto de visemas (fonemas visuales) que los representen. Algunos de los modelos que se consiguieron en la red representan visemas en inglés, pero lo que necesitamos es un conjunto de modelos que representen visemas en español. Hay algunos visemas en inglés que se parecen a los del español y se usaron sin ninguna modificación, para los visemas en español que no estaban representados por los visemas originales se generaron nuevos visemas combinando los originales con la técnica de metamorfosis ponderada que se explicó en la sección anterior.

A continuación se presenta la propuesta de correspondencia entre fonemas y visemas para el idioma Español que se utilizó en este trabajo, empezando por la proyección de los fonemas correspondientes a las vocales, que son los que tienen mayor contraste visual. La Figura 5-3 muestra los visemas de las vocales. Como se expuso en el Capítulo 3, el sonido de las vocales es dado por el grado de constricción de la cavidad oral, el cual varía en gran medida para cada vocal como se aprecia en la figura.

En cuanto a las consontantes, éstas se pueden agrupar en categorías de visemas de acuerdo a su punto de articulación. Por ejemplo, /b/, /m/ y /p/ son consonantes labiales, ya que

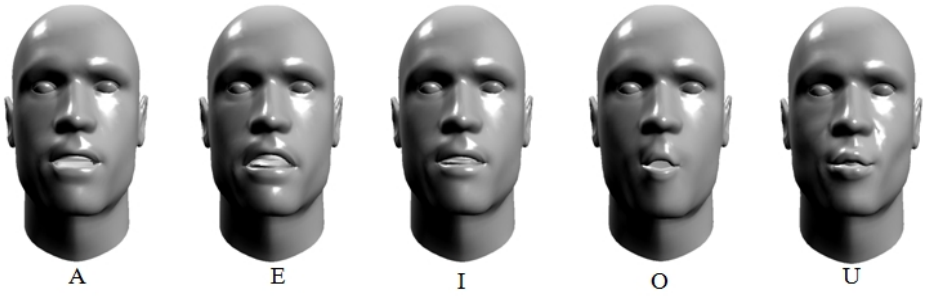

Figura 5-3: Visemas para los fonemas /a/, /e/, /i/, /o/, /u/.

son articuladas contra el labio superior, por lo que caen en la misma categoría de visema. Las consonantes /f/ y /v/ son labiodentales y comparten una categoría de visema, siendo ésta una de las más marcadas del conjunto. Entre las consonantes dentales encontramos a la /t/ y /d/. Los sonidos /ch/ y /x/ como en *Xicotencatl* pueden ser agrupados en una categoría, para estos sonidos se usó el visema correspondiente al sonido /sh/ en inglés. Aunque las consonantes alveolares /n/, /s/, /l/ y /r/, pertenecen a la misma clasificación en cuanto al punto de articulación, contrastan visualmente y deben ser puestas en distintas categorías de visema. Otros sonidos que sí pueden agruparse son /k/ y /x/ como en *México*, este último representa también a la *g* y *j*, todos estos sonidos son velares y su punto de articulación es el dorso de la lengua y el velo del paladar, por lo que es difícil identificarlos visualmente. La Figura 5-4 muestra los visemas que corresponden a las consonantes.

Como se puede observar en la Figura 5-4, los visemas no representan a todas las consonantes de nuestro abecedario, ya que distintas consonantes pueden compartir el mismo sonido y por consiguiente, la misma categoría de visema. Por ejemplo, el sonido de *k* en *Karla*, es el mismo que el de *c* en *casa*. En la siguiente sección se presenta el proceso de análisis fonológico, donde se obtiene la lista de visemas correspondiente a un texto en español.

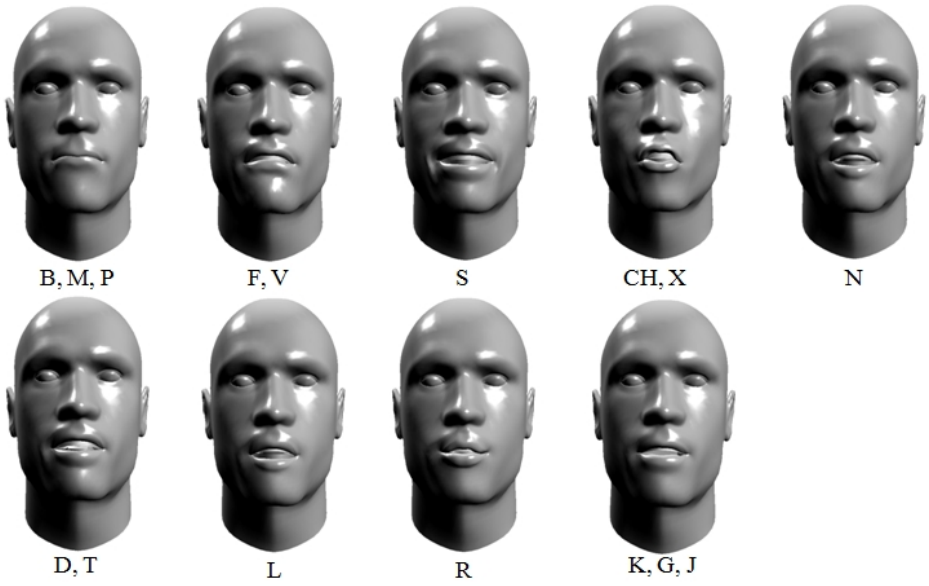

Figura 5-4: Visemas correspondientes a los sonidos de las consonantes. Estos sonidos están agrupados en categorías de visema.

5.5. Análisis Fonológico

Para poder obtener la lista de visemas que representen visualmente la pronunciación de un determinado texto, es necesario hacer un análisis fonológico del mismo y extraer los fonemas para poder proyectarlos a su conjunto de visemas. En la sección anterior, se muestra el conjunto de visemas usado en este trabajo, el cual no representa todas las consonantes de nuestro alfabeto, ya que los visemas representan sonidos, no letras, por lo que algunas consonantes se agrupan en categorías de visema. La Tabla 5-1 presenta la relación entre cada una de las letras del abecedario, sus fonemas representativos y sus categorías de visema (los nombres para las categorías de visema se toman de las Figuras 5-3 y 5-4 y su letra representativa se muestra subrayada).

Categoría de Visema	**Fonemas representativos**	**Letras correspondientes**
(A)	/a/	**a**, á
(E)	/e/	**e**, é
(I)	/i/	**i**, í, y, ll
(O)	/o/	**o**, ó
(U)	/u/, /w/	**u**, ú, ü, w
(B,M,P)	/b/, /m/, /p/	**b**, m, p
(F, V)	/f/, /v/	**f**, v
(S)	/s/, /θ/	**s**, z, c (ce, ci)
(CH,X)	/č/	**x**, ch, sh
(N)	/n/	**n**, ñ
(D,T)	/d/, /t/	**d**, t
(L)	/l/	**l**
(R)	/r/, /R/	**r**, rr
(K,G,J)	/k/, /x/	c (ca, co, cu), **k**, g, j, q (que, qui)

Tabla 5-1: Relación entre letras, fonemas y categorías de visema.

Cuando el programa analiza el texto de entrada, se encarga de sustituir letras individuales o segmentos de la palabra por su sonido representativo, esto es, por una letra que represente el fonema apropiado. Uno de los casos más sencillos de este proceso es la eliminación de la letra *h*, esta letra en español no suena y por lo tanto no tiene un fonema propio. Por ejemplo, para representar la pronunciación de la palabra *hola* se puede escribir simplemente como *ola*. Incluso en casos como *hueso* o *huevo*, donde se le podría asignar a la *h* el sonido de *güeso* o *güevo*, si se elimina la *h* se obtiene una pronunciación similar debido a las vocales que le siguen, por lo que no se consideran estos casos como especiales.

La letra *c* tiene asignados dos fonemas distintos, el fonema /k/ se da en las sílabas *ca*, *co* y *cu*, mientras que el fonema /s/ se da en las sílabas *ce* y *ci*, el programa simplemente sustituye estas sílabas por *ka*, *ko*, *ku*, *se* y *si* respectivamente. De esta forma, una palabra como *canción* quedaría como *kansión* en su representación de fonemas. Los acentos son un caso especial, ya que en la animación es difícil mostrar visualmente la acentuación, sin embargo, en el proceso de silabificación, que se discutirá después, son un elemento importante para separar diptongos y

lograr una adecuada silabificación. Al obtener los fonemas de vocales acentuadas, simplemente se sustituye por la vocal no acentuada correspondiente, tomando de nuevo como ejemplo la palabra *canción*, ésta se representa como *kansion*.

Hay sustituciones de letras que se hacen no de acuerdo al propio fonema, sino a la categoría de visema a la que pertenecen. Por ejemplo, de la Figura 5-4 se puede observar que las letras *b*, *m* y *p* pertenecen a la misma categoría de visema, por lo que las letras *m* y *p* se sutituyen por la letra *b* (por ser ésta la primera de las tres en orden alfabético, pero otra sustitución entre ellas también es válida). Así, la palabra *prima* se representa como *briba*. Otras sustituciones similares son las de *t* por *d*, *v* por *f* y *z* por *s*, con las cuales palabras como *todo*, *voto* y *zapato* se representan como *dodo*, *fodo* y *sabado*. De estos ejemplos, se puede notar que usando esta aproximación se tienen las mismas secuencias de visemas para algunas palabras, incluso si éstas difieren totalmente, como es el caso de *zapato* y *sábado*, que a la hora de sustituir las letras por sus fonemas y categorías de visemas correspondientes, comparten la misma representación, a saber, *sabado*.

Existen segmentos dentro de las palabras que corresponden a un único sonido en español, como es el caso de *ch* y *ll*, que no se toman como dos consonantes sino como una sola. En el caso de *ch* se toma como el sonido [x] en *Xicotencatl*. Cabe señalar también, que el programa no distingue entre las distintas pronunciaciones de la letra *x*, por ejemplo, en *México* y *Tlaxcala* a la *x* se le asigna el visema correspondiente a *ch*, hacer tal distinción es difícil debido a que no hay reglas específicas para su pronunciación. En cuanto a *ll*, este segmento se sustituye por *y* como en la palabra *yeso*, que a su vez se sustituye por la vocal *i* para representarla visualmente. Aunque tal vez esta sustitución no sea la adecuada, si uno pronuncia *yeso* como *ieso*, o *lluvia* como *iuvia*, la diferencia no es mucha y se logra entender lo que se está diciendo. Lo mismo ocurre con la conjunción *y*, como en *mamá y papá*, que se representa como *baba i baba*, de acuerdo a las sustituciones que se han presentado, note de nuevo como la palabra *mamá* y la palabra *papá* comparten la misma representación.

Otros segmentos que requieren atención especial son las sílabas *gue*, *gui*, *que* y *qui*, que aunque están compuestas por tres letras, sus sonidos corresponden unicamente a dos, a saber, *ke* y *ki*. Obviamente, el sonido de *gue* es distinto al sonido de *que*, sin embargo, ambos son velares y no son visualmente contrastantes por lo que se han ubicado en la misma categoría de

visema, junto con la *j*. De esta forma, palabras como *gato*, *guitarra*, *quien*, *quemadura* y *jilguero* se representan visualmente como *kado*, *kidarra*, *kien*, *kebadura* y *kilkero*. Por último, aunque no es muy común en español encontrar palabras con *w*, ésta se sustituye por la vocal *u*, por ejemplo, la palabra *Wisconsin*, se representa visualmente como *Uiskonsin*.

La Figura 5-5 muestra el proceso de obtención de los visemas para la palabra *zapato*, como se explicó anteriormente primero se sustituyen las letras por su sonidos representativos y categorías de visema. La *z* se sustituye por *s*, ya que es el sonido que la representa (aunque en España es más marcado el contraste entre /s/ y /θ/, en México no suele haber tal distinción en la pronunciación), la *p* se sustituye por *b*, debido a que pertenece a la misma categoría de visema que esta última, lo mismo sucede con la *t* que se sustituye por *d*. Después de realizar estas sustituciones, cada una de las letras se proyecta a su correspondiente visema y se logra animar la pronunciación de la palabra.

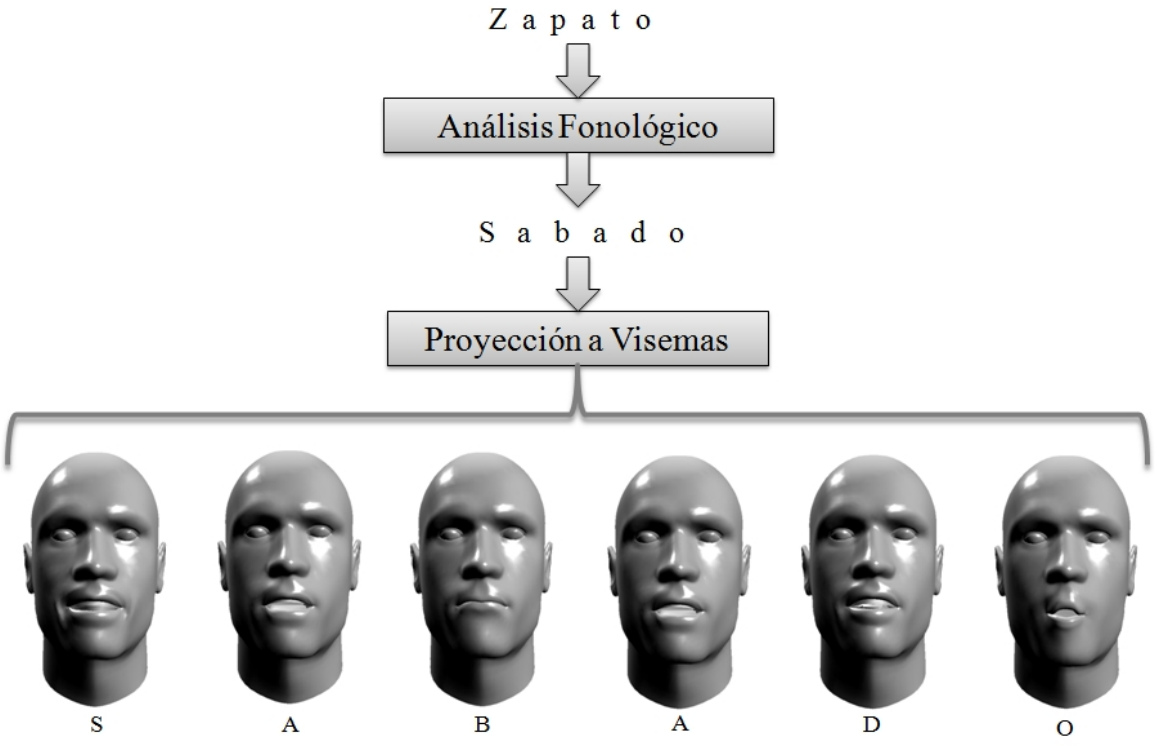

Figura 5-5: Proyección de la palabra *zapato* a su conjunto de visemas representativo.

5.6. Adición de Expresiones Faciales

Hasta ahora hemos visto en qué consiste el proceso de animación de la pronunciación de un texto en español, pero se ha hecho a un lado la expresividad del rostro. Como se muestra en la Figura 5-1, se cuenta con varios modelos 3D para representar emociones, sin embargo, obtener información respecto al estado de ánimo del personaje virtual mediante un análisis semántico del texto, sería una tarea complicada y se saldría del alcance del presente trabajo, por lo que se implementó un sistema tipo *script* para añadir expresividad al rostro. Se reservaron secuencias de caracteres especiales para delimitar un fragmento de texto sobre el cual actúa un modificador de expresión dado. Por ejemplo, la secuencia =) corresponde al gesto de sonreír y todo el texto que se encuentre entre dos ocurrencias de esta secuencia, será pronunciado junto con una sonrisa en el rostro. La Figura 5-6 muestra las secuencias de caracteres para las principales expresiones faciales, aunque no son las únicas.

Secuencia de caracteres	Icono	Expresión
=[		Enojo 
={		Disgusto 
=$		Miedo 
=(		Tristeza 
=)		Sonrisa 
=*		Sorpresa 

Figura 5-6: Algunas expresiones faciales y sus correspondientes secuencias de caracteres e iconos.

Se puede observar que además de tener una secuencia de caracteres asignada, cada expresión tiene asociado un icono (comúnmente conocido como *emoticon*). La interfaz del programa final contiene una serie de botones para añadir estos iconos al texto, como se hace en los populares programas de mensajería instantánea. La primer ocurrencia de un icono (o su correspondiente secuencia de caracteres) determina el punto de inicio de aplicación del modificador de expresión, todo el texto que le siga será pronunciado con esa expresión en el rostro, el modificador actúa hasta encontrar otra ocurrencia del *mismo* icono. Las expresiones se pueden combinar como desee el usuario, sin embargo, combinar muchas expresiones puede deformar el rostro y el resultado no sería muy agradable. La Figura 5-7 muestra la animación de la pronunciación de la palabra *feliz* con un modificador de sonrisa aplicado.

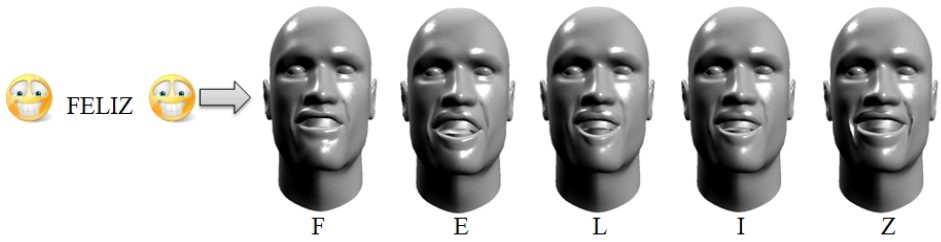

Figura 5-7: Pronunciación de la palabra *feliz* con una expresión de felicidad.

La combinación de expresiones faciales y de visemas se hace utilizando la técnica de combinación ponderada descrita anteriormente, por lo que los desplazamientos de los vértices se ven afectados tanto por el modelo del visema como por el de la expresión. Esta combinación de expresiones y visemas hace que se pierda un poco la claridad de los visemas y puede llegar a provocar que no se entienda la palabra que se está pronunciando. Este efecto es contraproducente, ya que el objetivo principal del trabajo es que se pueda leer los labios del modelo y entender lo que está diciendo. A pesar de esto, la adición de expresividad al rostro se logra de manera interactiva y provee al usuario de una herramienta más para modificar la animación.

5.7. Los Tiempos de la Animación

Ya hemos visto cómo se obtienen los visemas a partir de un texto y cómo se añade expresividad al rostro, con esto se tiene la información de los fotogramas clave de la animación,

pero para lograr una transición suave entre un gesto del rostro y otro, es necesario asignar tiempos adecuados a estas transiciones. Para esta asignación se tienen dos aproximaciones: la primera consiste en asignar el mismo tiempo a cada sílaba (definido por el usuario), mientras en la segunda se trata de sincronizar la animación con la pista de audio correspondiente a la pronunciación del texto de entrada. A continuación, se presentan de manera detallada ambas aproximaciones y su implementación.

5.7.1. Tiempo por Sílaba

Para lograr una animación más realista, es necesario analizar la forma en que pronunciamos las palabras al hablar. La experiencia nos dice que, cuando intentamos separar la pronunciación de las palabras en periodos iguales de tiempo, la unidad fonológica más grande para hacerlo es la sílaba. Por lo cual, se ha considerado apropiado asignar los tiempos por sílabas, en lugar de hacerlo por cada visema, hacerlo de este modo implicaría por ejemplo, que al pronunciar la palabra *mamá*, el visema correspondiente al fonema /m/ tendría la misma duración que el del fonema /a/, dando como resultado una animación lenta y poco natural.

El primer paso para hacer una asignación de tiempo por sílaba es, naturalmente, el proceso de silabificación. En el Capítulo 3, se presentan una serie de reglas para la silabificación del español, ahora veremos cómo se implementan estas reglas en el sistema. Las reglas de silabificación que se utilizan son cuatro:

- Identificación de núcleos.
- Regla CV.
- Regla de Formación de Grupos de Ataque.
- Reglas de Adjunción de Codas.

La identificación de núcleos es una tarea sencilla, que se logra buscando la ocurrencia de una vocal en la palabra (ya que en español las vocales son las únicas que constituyen un núcleo silábico), una vez que la vocal se ha encontrado se busca la segunda ocurrencia de vocal en la palabra, con el objetivo de marcar los límites de la sílaba. Puede ocurrir el caso de que la palabra sea monosílaba y no encontrarse una segunda vocal en la palabra, si este es el caso ya

no es necesario procesar la palabra, ya que se ha encontrado la única sílaba en ella. En caso de encontrarse otra vocal, se pueden dar los siguientes casos: la segunda vocal forma un diptongo con la primera, por lo que correspondería a la misma sílaba, o bien, forma hiato con la primera o simplemente no está junto a ella. En el caso de formarse diptongo, se debe ignorar la segunda vocal y buscar una tercera para delimitar a la sílaba. En cualquier otro caso, se cuenta con dos índices que indican la posición del núcleo de la primer sílaba y el núcleo de la segunda.

El siguiente paso es aplicar la regla CV, esta regla silabifica la consonante anterior con el núcleo de la sílaba. Para implementar esta regla, simplemente se debe recorrer el índice del núcleo un lugar atrás si la letra anterior es una consonante; esta regla se aplica a los dos núcleos que se han obtenido. Posteriormente, se procede a maximizar los grupos de ataque, para ello se deben tener identificados los grupos de ataque permitidos en español, que son los formados por oclusiva o /f/ seguida de líquida. Los consonantes oclusivas son: *p*, *t*, *k*, *c*, *b*, *d*, *g*, *s*, *x*, *y*, se añade también *f*. Las líquidas son: *r* y *l*. El grupo de ataque que no se acepta es el formado por *dl* y los grupos *ch*, *rr* y *ll* se toman como una sola consonante. La maximización de grupos de ataque, se logra simplemente recorriendo el índice un lugar hacia atrás, cuando la consonante anterior forma un grupo de ataque admisible con la consonante actual.

Una vez que se han aplicado estas reglas, el primer índice indica donde empieza la primer sílaba y el segundo índice donde empieza la segunda, así que se puede separar el fragmento de la palabra delimitado por estos índices para obtener la primer sílaba. Este proceso se repite hasta que se haya reducido la longitud de la palabra a cero.

Para lograr la unión de los diptongos, o en su caso, la separación de los hiatos, las vocales se clasifican en fuertes y débiles. Las vocales débiles son: *i* y *u*, también se incluye a la *ü*. Las vocales fuertes son: *a*, *e* y *o*, y se incluyen a las vocales acentuadas *á*, *é*, *í*, *ó* y *ú*. Con esta clasificación, simplemente se aplica la regla que dice que siempre que hay una vocal débil existe diptongo y se logra una silabificación aceptable. Debe recordarse, como se explicó en el Capítulo 3, que hay excepciones que desgraciadamente no son consideradas por el algoritmo, por ejemplo, la palabra *huida* debería silabificarse como *hu-i-da*, pero al aplicar la regla de las vocales débiles se forma un diptongo y se silabifica como *hui-da*, lo que es correcto por ejemplo en *cui-da*. En el Capítulo 6, se discutirá más a fondo el desempeño del algoritmo de silabificación, por ahora basta con decir que se logra una silabificación aceptable en la mayoría de los casos.

Ya que se tienen las sílabas que forman la palabra, se les asigna el mismo tiempo a cada una de ellas, el cual será divido entre los fonemas que la forman. Siguiendo las reglas de silabificación estudiadas previamente, que dicen que los grupos de ataque aceptables están formados por oclusiva o /f/ seguida de líquida, que una vocal conforma el núcleo silábico y que la coda puede estar compuesta por una deslizada, una consonante, una deslizada seguida de consonante, o dos consonantes, se puede asumir que una sílaba puede tener a lo más cinco letras y ser de la forma CCVCC o CCVVC (con sus respectivos subconjuntos). Así, se divide el tiempo asignado a la sílaba de la siguiente manera: si la sílaba tiene ataque y coda, 40 % del tiempo se le asigna al núcleo, 30 % al ataque y 30 % a la coda; si la sílaba contiene ataque pero no coda, 60 % se le asigna al núcleo y 40 % al ataque; si la sílaba no contiene ataque pero sí coda, 60 % se le asigna al núcleo y 40 % a la coda; por último, si la sílaba solamente tiene núcleo, se le asigna el 100 % del tiempo. Ahora bien, dentro de la sílaba existe una escala de sonoridad, los fonemas que están más cerca del núcleo tienen más sonoridad que los más alejados, de tal forma que, el tiempo asignado al ataque y a la coda será divido entre los fonemas que los forman bajo el siguiente criterio: suponiendo que el ataque y la coda tienen a lo más dos letras, se le asignará 60 % del tiempo a la letra más cercana al núcleo y 40 % a la más alejada; en caso de solamente contar con una letra, se le asignará el 100 % del tiempo a ésta. De esta forma, se logra simular visualmente la escala de sonoridad dentro de la sílaba para lograr una animación más realista.

5.7.2. Sincronización con Audio

Como se plantea en los objetivos específicos del sistema, se pretende sincronizar la animación con el audio correspondiente a la pronunciación del texto de entrada. Para ello, se experimentó con dos técnicas de sincronización: la primera consiste en utilizar un sintetizador de texto a voz para que el sistema genere tanto la animación como el archivo de audio; la segunda se vale de un archivo de audio existente para obtener los tiempos de la animación. En esta sección se discuten las dos técnicas y sus detalles de implementación.

Sintetizador de Texto a Voz

En el mercado existen muchos sintetizadores de texto a voz y lo que hacen principalmente es generar un archivo de audio a partir de un archivo de texto, el archivo de audio corresponde a la

pronunciación de las palabras del texto por una voz artificial, la calidad de la voz artificial varía desde una voz robotizada a algunas muy naturales. El sintetizador con el que se experimentó fue Festival, éste es un sintetizador multilingüe desarrollado por The Centre for Speech Technology Research (CSTR), en The University of Edinburg y es distribuido como software libre. Además, se trabajó con la extensión OGISpanish para Festival desarrollada por el equipo OGI y Alejandro Barbosa de la Universidad de las Américas Puebla, esta extensión provee voces de hombre y de mujer mexicanos para el español.

Aunque el desempeño del sintetizador es bueno, se tuvieron algunos problemas al tratar de conectarlo con el sistema principal. El principal problema fue que Festival funciona en sistemas operativos UNIX y en ese momento, el sistema estaba siendo desarrollado bajo Windows, lo que implicaba mudar toda la aplicación a un sistema operativo como LINUX, solamente para poder utilizar el sintetizador. Afortunadamente, el sistema fue desarrollado en Java lo que lo hace portable, sin embargo, para lograr la sincronización entre el archivo de audio y la animación era necesario adentrarse más a la programación interna de Festival para controlar los tiempos del sintetizador, además de que, el audio resultante presentaba una voz muy artificial y no se escuchaba muy bien. Fue en ese momento cuando se decidió experimentar con otra aproximación para la sincronización de la animación y el audio.

Tiempos a partir de un Archivo de Audio

Esta técnica consiste en obtener los tiempos para cada fonema a partir de un archivo de audio, en la literatura hay mucha información al respecto y sobre su implementación. En este trabajo, se usó la aplicación SAPI 5.1 Lipsync que provee Annosoft como software libre. Este programa es una implementación de la interfaz de programación Microsoft Speech API 5.1 y se encarga de realizar un alineamiento de fonemas en el tiempo a partir de un archivo de audio en el formato Microsoft RIFF Wave (.wav). Las entradas del programa son un archivo .wav y un archivo de texto con una transcripción opcional. El modo que usa el archivo de audio unicamente es llamado *textless lipsync* y el modo que usa también la transcripción es llamado *text based lipsync*. La salida del sistema es una lista de tiempos de fonemas y tiempos de palabra delimitados por caracteres de fin de línea, producidos por SAPI. La Figura 5-8 muestra el formato de la salida del programa SAPI 5.1 Lipsync. El marcador **phn** describe un evento de

fonema, primero se describe el tiempo de inicio del fonema en milisegundos, después el tiempo de finalización del fonema también en milisegundos, el tercer valor es llamado *morph-value* y es generado por el SDK de Annosoft, su valor es 0 si representa un silencio o 75 en caso contrario. El cuarto valor de la etiqueta **phn** describe el nombre de etiqueta del fonema, conforme a las etiquetas de fonemas de Annosoft [59].

```
audio C:\Users\Hector\Pictures\AUD000004.wav
phn 0 1500 75 x
word 1500 1640 que
phn 1500 1552 75 d
phn 1552 1640 75 AE
word 1640 2100 bonito
phn 1640 1689 75 CH
phn 1689 1787 75 AO
phn 1787 1837 75 IH
phn 1837 1919 75 AH
phn 1919 2001 75 AH
phn 2001 2100 75 AO
word 2100 2250 dia
phn 2100 2157 75 AY
phn 2157 2192 75 n
phn 2192 2249 75 AH
%%-begin-anno-text-%%
que bonito dia
%%-end-anno-text-%%
```

Figura 5-8: Salida del programa SAPI 5.1 Lipsync. Se obtienen los tiempos a partir de un archivo de audio para el texto *que bonito día*.

La salida del programa contiene otros marcadores, como el marcador **word** que describe un evento de reconocimiento de palabra, con sus tiempos de inicio y término correspondientes. Sin embargo, el único marcador que nos interesa es el marcador **phn**, ya que a partir de él se pueden obtener los tiempos para los visemas. Este marcador indica el tiempo en que inicia la pronunciación de un fonema y el tiempo en que finaliza, restando estos valores se obtiene la duración en milisegundos del fonema, que después se le asigna al visema correspondiente. La aplicación fue desarrollada para el idioma inglés, por lo que se debe tener cuidado, sobre todo al escribir la transcripción, de incluir solamente las representaciones en fonemas de las palabras, por ejemplo, escribir *ola* en lugar de *hola*, o *kien* en lugar de *quién*, ya que el programa busca la correspondencia entre la letra y la pronunciación del fonema en el archivo de audio.

La idea original era llamar desde el programa principal a la aplicación SAPI Lipsync 5.1, pasándole como argumentos el nombre del archivo de audio y la transcripción correspondiente, para después leer la salida y obtener los tiempos para los visemas. Sin embargo, se presentaron

algunos problemas de sincronización y no fue posible hacerlo de esta manera, así que se le incluyó al sistema la capacidad de leer un archivo en el formato de salida del programa SAPI Lipsync 5.1, para obtener los tiempos de la animación. Con esto, se logra generar una animación que puede ser reproducida junto con el archivo de audio original y presentar una buena sincronización. En la Figura 5-9 se muestra un diagrama de bloques que representa de manera resumida el procedimiento para generar la animación del rostro. En el Capítulo 6, se discute más acerca del desempeño de la animación y la opción para grabarla en video y reproducirla después.

5.8. Funcionamiento del Sistema

En esta sección se describe el funcionamiento del programa final, se exponen principalmente, las funciones de la interfaz gráfica de usuario. Las secciones anteriores, están enfocadas a describir el funcionamiento interno del sistema, se discuten los algoritmos y técnicas empleadas para lograr la animación del rostro. Ahora pasaremos a un estudio del sistema desde el punto de vista del usuario final. La Figura 5-10 es un diagrama del funcionamiento general del sistema, en ella se muestra la interfaz gráfica de usuario, así como también las clases y métodos responsables de la interacción con el usuario. Además, se muestra la dinámica entre estas clases para ilustrar el funcionamiento interno del sistema.

Al iniciar el programa, se construye la escena tridimensional por medio de un objeto de la clase **WrapLoader3D**, éste agrega la luz ambiental, las luces direccionales y el fondo de la escena. Una instancia de la clase **PropManager** se encarga de cargar los modelos 3D y guardarlos como objetos de la clase **MyMorph**, con la cual se realiza la metamorfosis y combinación de los modelos.

Un objeto de la clase **MorphBehavior** es el encargado de la animación, dentro de su método **processStimulus** se realiza en cada fotograma la interpolación entre modelos, modificando el valor del parámetro t de la ecuación de interpolación lineal y pasándolo como argumento al método **morphShape** de las instancias de la clase **MyMorph**.

Cuando el usuario presiona el botón **OK**, el texto que se ha escrito en el panel se pasa como argumento el método **speak** de la clase **MorphBehavior**, éste a su vez, lo pasa el método **convertToVisemes** de la clase **VisemeConverter**. La instancia de esta clase procesa el texto

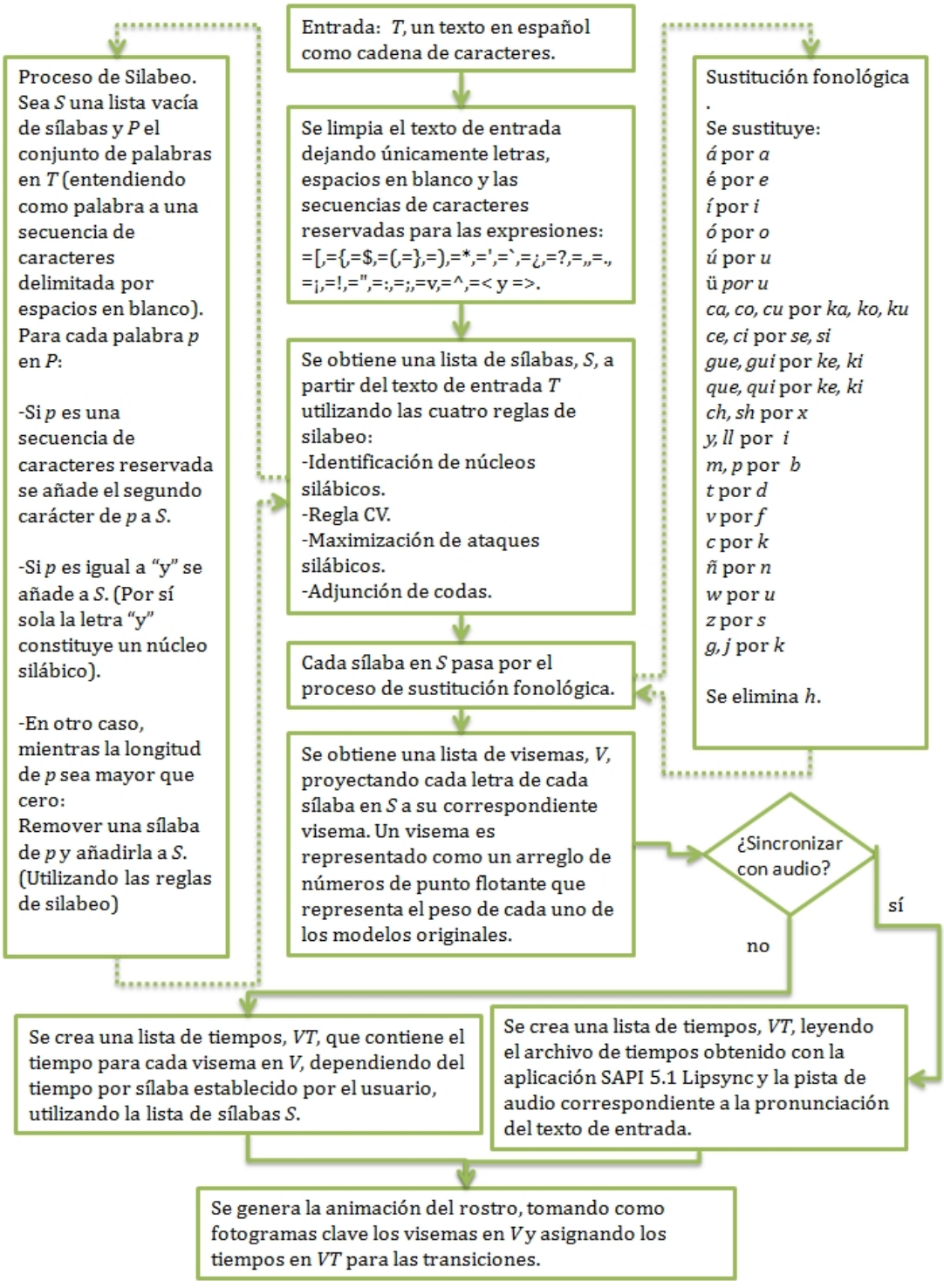

Figura 5-9: Diagrama de bloques del procedimiento de animación del rostro.

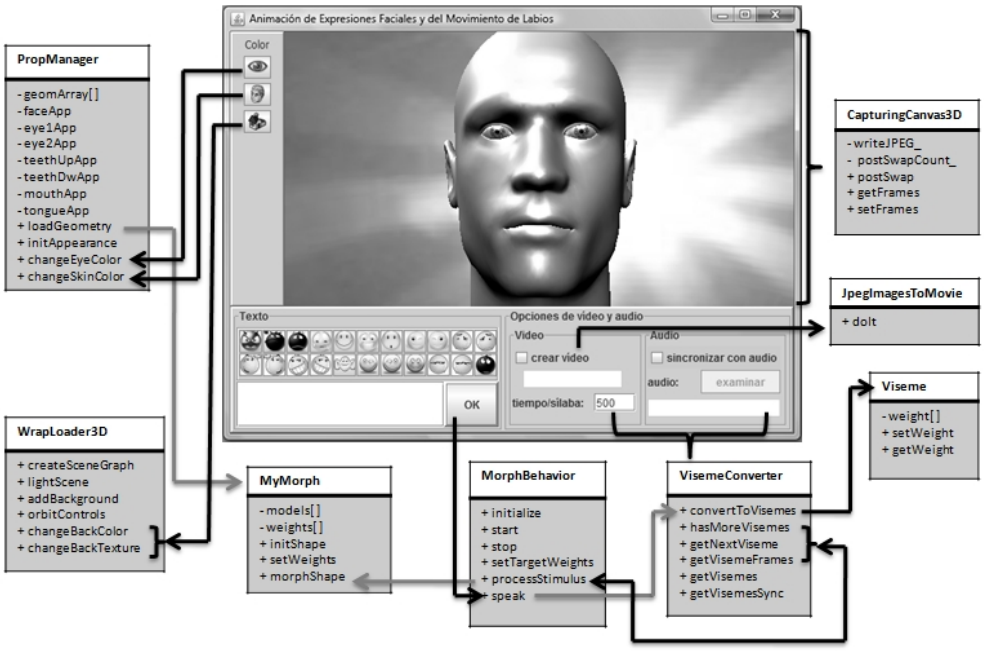

Figura 5-10: Diagrama del funcionamiento general del sistema.

de entrada, generando la lista de visemas correspondiente. También, se encarga de asignar los tiempos a cada uno de estos visemas, ya sea por medio de la silabificación o la sincronización con un archivo de audio. La animación se desarrolla mientras haya visemas que procesar, para ello existen los métodos **hasMoreVisemes** y **getNextViseme**, que son llamados dentro del método **processStimulus** del objeto **MorphBehavior**. El tiempo de la animación se mide en fotogramas por segundo, así que el tiempo asignado a cada visema es medido en fotogramas, con la ayuda del método **getVisemeFrames** se obtiene el tiempo para el visema que es procesado.

La interfaz de usuario ofrece además, algunas herramientas para modificar la apariencia de los objetos y la escena 3D. La primera de estas herramientas, permite al usuario modificar el color de ojos del modelo; para lograr esto, se aplica una técnica de falso color sobre una imagen de un iris en escala de grises, tomando como referencia el color que haya seleccionado el usuario. Se utilza de nuevo la interpolación lineal para el falso color, pero esta vez se hace por partes, primero se interpola entre el color negro y el nuevo color, después, entre el nuevo color y el color blanco, esto se hace así para no modificar regiones de la imagen que no corresponden al iris. Finalmente, la imagen generada por falso color se aplica como textura a los objetos que

forman los ojos.

El color de la piel del modelo también puede modificarse, para ello, simplemente se le asigna el color que seleccione el usuario al color *difuso* del material para la piel del modelo. El color difuso del material define el color de éste cuando es iluminado por una fuente de luz. En general, un material tiene otros componentes de color, como el color *ambiente* que define cuanta luz ambiental es reflejada por el material y el color de *emisión*, que define el color de la luz que emite el material (si es que emite luz).

Por último, el usuario puede modificar también el color de fondo o, si lo desea, establecer una imagen como fondo. El fondo de la escena está compuesto por una esfera que envuelve a todos los objetos en ella, cuando se modifica el color de fondo en realidad se modifica el color de emisión de la esfera, a diferencia de los objetos 3D en la escena, la esfera del fondo no es afectada por las luces que existen, por lo que la esfera emite luz por sí misma y se modifica el color de esta luz. Para aplicar una imagen al fondo, se añade una textura a la esfera que lo forma, pero antes se invierten los vectores normales de la esfera para que la textura se pueda ver, ya que la escena está dentro de ella.

Por otra parte, para poder guardar en video la animación del rostro, se utilizan las clases **CapturingCanvas3D**[2] y **JpegImagesToMovie**[3]. La instancia de la clase **CapturingCanvas3D**, captura una imagen de la escena 3D y la guarda en formato jpeg para cada fotograma de la animación. Cuando la animación termina, una instancia de la clase **JpegImagesToMovie**, toma las imágenes jpeg que se guardaron y las utiliza para producir un video en el formato .mov. Este video puede ser reproducido después por un reproductor externo.

En este capítulo se expuso el proceso de diseño e implementación del sistema de animación facial, analizando cada uno de los procesos que lo componen y sus detalles de implementación. En el siguiente capítulo, se evalúa el desempeño del programa, presentando las pruebas que se realizaron y los resultados obtenidos.

[2] **La clase CapturingCanvas3D fue desarrollada por Peter Z. Kunszt en la Universidad Johns Hopkins (Johns Hopkins University). Se encarga de capturar una imagen fija en formato jpeg.**

[3] **La clase JpegImagesToMovie es propiedad de Sun Microsystems y es distribuida libremente.**

Capítulo 6

Pruebas y Resultados

En este capítulo se presenta un estudio del desempeño del sistema, se evalúa la funcionalidad de los procesos que lo forman y las aplicaciones que puede tener. El estudio es dividido de acuerdo a los módulos que componen al sistema y se exponen las pruebas y resultados de cada uno de ellos. Primero, se evalúa la técnica de metamorfosis ponderada para la combinación de expresiones faciales y visemas. Después, se muestran los resultados del analizador fonológico, desde el proceso de silabificación hasta la proyección a visemas. Posteriormente, se evalúa la animación en sí, en particular, su desempeño en relación al tiempo de procesamiento. Y por último, se presenta el plan de trabajo a realizar con la Escuela de Educación Especial Jean Piaget de la ciudad de Puebla, junto con la evaluación preliminar del uso del sistema como herramienta auxiliar en la terapia del lenguaje.

6.1. Técnica de Metamorfosis Ponderada

En este trabajo se implementó la técnica de metamorfosis ponderada sobre la plataforma de desarrollo de Java 3D para la combinación de modelos tridimensionales. Los resultados obtenidos con esta técnica son satisfactorios en comparación con la técnica de metamorfosis ponderada de la clase **Morph** de Java3D, ya que no se tienen restricciones en cuanto a la forma de combinar los modelos y los pesos se pueden asignar libremente a cada modelo sin que estén restringidos a un determinado rango. El usuario es entonces el responsable de combinar los modelos de manera adecuada para no deformar demasiado el rostro. La Figura 6-1 muestra un

ejemplo de modelos obtenidos con la metamorfosis ponderada, empezando por el rostro neutral y combinando cada vez más modelos. Como se puede observar en la Figura 6-1f, el modelo final es una combinación de las expresiones faciales aplicadas y el visema correspondiente a la letra *a*.

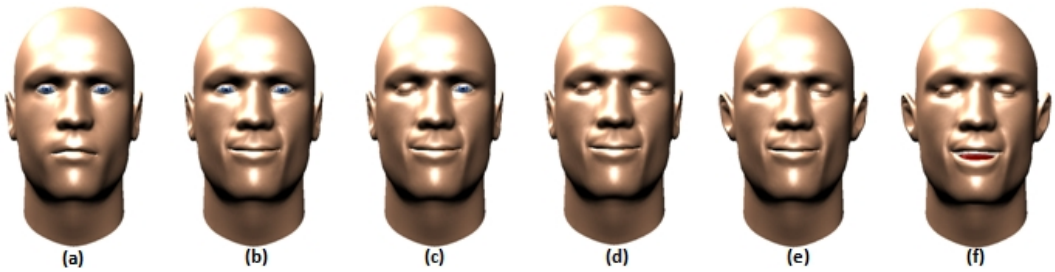

Figura 6-1: Combinación de modelos 3D usando metamorfosis ponderada. (a) Modelo base, (b) más la expresión de sonrisa, (c) más ojo derecho cerrado, (d) más ojo izquierdo cerrado, (e) más orejas hacia afuera, (f) más el visema *a*.

6.2. Resultados del Analizador Fonológico

Una parte esencial del sistema es el analizador fonológico, éste se encarga de descomponer el texto de entrada en visemas para generar la animación. Este proceso se realiza en tres etapas distintas: primero se silabifica el texto para calcular los tiempos de los visemas (si se sincroniza con un archivo de audio la silabificación no es necesaria), después se representa el texto en fonemas, es decir, se sustituyen las palabras por su representación fonológica y finalmente, se proyectan estos fonemas a sus visemas correspondientes. A continuación se presentan los resultados correspondientes a la implementación de cada una de estas etapas.

6.2.1. Proceso de Silabificación

En este trabajo, se propone un algoritmo de silabificación basado en un estudio serio de fonología de nuestro idioma[51], presentado en el Capítulo 3. Este algoritmo consta de cuatro pasos principales: identificación de núcleos silábicos, silabificación del núcleo con la vocal anterior, maximización de ataques silábicos y adjunción de codas. A continuación se muestra un fragmento de texto silabificado con este algoritmo.

Cier-to hom-bre, que ha-bí-a com-pra-do u-na va-ca mag-ní-fi-ca, so-ñó la mis-ma no-che que cre-cí-an a-las so-bre la es-pal-da del a-ni-mal, y que és-te se mar-cha-ba vo-lan-do. Con-si-de-ran-do es-to un pre-sa-gio de in-for-tu-nio in-mi-nen-te, lle-vó la va-ca al mer-ca-do nue-va-men-te, y la ven-dió con gran pér-di-da. En-vol-vien-do en un pa-ño la pla-ta que re-ci-bió, la e-chó ***so-bre*** *su es-pal-da, y a mi-tad del ca-mi-no a su ca-sa, vio a un hal-cón co-mien-do par-te de u-na lie-bre. A-cer-cán-do-se al a-ve, des-cu-brió que e-ra bas-tan-te man-sa, de ma-ne-ra que le a-tó u-na pa-ta a u-na de las es-qui-nas con el pa-ño en que es-ta-ba su di-ne-ro. El hal-cón* ***a-le-te-a-ba*** *mu-cho, tra-tan-do de es-ca-par, y tras un ra-to, al a-flo-jar-se* ***mo-men-tá-ne-a-men-te*** *la ma-no del hom-bre, vo-ló con to-do y el tra-po y el di-ne-ro. "Fue el des-ti-no", di-jo el hom-bre ca-da vez que con-tó la his-to-ria; ig-no-ran-te de que, pri-me-ro, no de-be te-ner-se fe en los sue-ños; y, se-gun-do, de que la gen-te no de-be re-co-ger co-sas que ve al la-do del ca-mi-no. Los cua-drú-pe-dos ge-ne-ral-men-te no vue-lan.*

Este fragmento se tomó de un artículo de Heriberto Cuayáhuitl[60], quien propone otro algoritmo de silabificación para el español. Se han resaltado tres palabras en negritas, la primera palabra *sobre*, es silabificada de manera diferente por el algoritmo de Cuayáhuitl; su algoritmo separa los prefijos como *sobre*, *archi*, *trans*, *ante* y otros más, tomándolos como una sola sílaba, sin embargo, esto tiene como resultado que palabras como *sobre* no se silabifiquen, o palabras como *extraer* se silabifiquen como *extra-er*, lo cual no es correcto. Aunque esto podría solucionar problemas como por ejemplo, con la palabra *sublingual*, la cual nuestro sistema silabifica como *su-blin-gual* en lugar de *sub-lin-gual*, es difícil generalizar la aplicación de la técnica de los prefijos, ya que por ejemplo, si se tomara el prefijo *sub* siempre como una sílaba aparte, la palabra *subliminal* sería silabificada como *sub-li-mi-nal*. Con el algoritmo propuesto aquí esto no sucede, además de que consta de menos pasos que el de Cuayáhuitl. Las otras dos palabras resaltadas, *aleteaba* y *momentáneamente*, se silabifican igual que como lo hizo Cuayáhuitl, pero la silabificación no es la idónea, ya que podrían silabificarse como *a-le-tea-ba* y *mo-men-tá-nea-men-te*, debido a un fenómeno conocido como *sinéresis*, que ocurre en el habla rápida y reduce los hiatos dentro de la palabra[51].

En general, el desempeño del algoritmo es bueno y son las excepciones a las reglas de silabificación en español las que provocan los errores en el proceso. En el presente trabajo, la silabificación se utiliza para asignar tiempos a los visemas y el algoritmo que se propone aquí

funciona bien para este propósito.

6.2.2. Sustitución Fonológica

Esta parte del análisis fonológico del texto es relativamente sencilla, simplemente se sustituyen algunas letras por la letra que mejor represente a su fonema correspondiente. Existe un caso especial que puede significar un defecto en el sistema y es el de la letra *x*. Esta letra en el español de México puede tener varias pronunciaciones y desgraciadamente, no existen reglas definidas para saber cuándo se pronuncia de una forma y cuándo de otra. Por ejemplo, en la palabra *Tlaxcala* se pronuncia parecida a la *s*, mientras que en la palabra *México* como *g*. El sistema no diferencia entre estas pronunciaciones y todas las ocurrencias de la letra *x* se asume que se pronuncian como *ch*.

Otro aspecto del sistema que puede considerarse débil, es el hecho de que se ignora la acentuación de las palabras. Visualmente es difícil mostrar esta acentuación, sin embargo, se intenta hacerlo asignando tiempos diferentes a cada visema dentro de una sílaba, tomando como referencia una escala de sonoridad donde los fonemas más alejados del núcleo tienen menor sonoridad que los más próximos. En base a los resultados obtenidos, se puede concluir que la sustitución fonológica que se realiza, aunque tenga detalles de implementación que se pueden mejorar, es adecuada para generar la animación del rostro y comunicar un mensaje a través de la lectura de labios.

6.2.3. Proyección a Visemas

Según la aproximación que se siguió en este trabajo, distintos fonemas pueden caer en la misma categoría de visema, esto tiene como resultado que palabras muy diferentes puedan generar la misma animación. Esto sucede como ya hemos visto, en palabras como *zapato* y *sábado*, o *mamá* y *papá*, que visualmente son iguales aunque en realidad son palabras muy diferentes. Esto lleva a la conclusión, de que en la lectura de labios están involucrados otros factores, como el contexto y el conocimiento del tema por parte del que intenta leer los labios. En la Figura 6-2, se muestra la animación de la frase *el sábado veo a mi papá*, sin embargo, esta misma animación correspondería a la frase *el zapato feo a mi mamá*, debido a la sustitución fonológica y la proyección a visemas. La lectura de labios no es sencilla y se requiere de mucha

habilidad para entender lo que alguien está diciendo de esta manera, pero el contexto en el que se desarrolle una conversación puede ayudar a deshacer la ambigüedad visual que existe entre varios fonemas.

Figura 6-2: Animación de la pronunciación de la frase *el sábado veo a mi papá*, que podría corresponder también a la frase *el zapato feo a mi mamá*.

6.3. Tiempos de la Animación

Uno de los objetivos de este trabajo es que la animación del rostro se genere "en tiempo real", es decir, que no sea necesario un procesamiento previo ni posterior del texto, o de la información de los fotogramas clave de la animación. Este objetivo se cubrió parcialmente, la animación se genera sin ningún tipo de procesamiento adicional, más que el del análisis del texto y la interpolación entre visemas. Sin embargo, la animación tarda un poco más del tiempo establecido por el usuario, o del tiempo de sincronización con audio. Por ejemplo, si se genera la animación de la palabra *hola* y se asigna un tiempo por sílaba de 500ms, se esperaría que la duración de la animación fuera de un segundo, ya que son dos sílabas, pero en general, la animación tarda más de lo esperado. Esto se debe al tiempo de procesamiento que se requiere para realizar la interpolación entre los visemas, recordemos que esta interpolación se hace dentro del método **processStimulus()** de la clase **MorphBehavior**, que procesa un fotograma cada milisegundo (este tiempo se estimó haciendo pruebas con el sistema), pero al realizar la interpolación, la computadora consume tiempo, llegando a alcanzar medio milisegundo de procesamiento, lo que no solamente retarda el tiempo de procesamiento del siguiente fotograma, sino que de hecho provoca que no se procesen los fotogramas cada milisegundo, sino mucho tiempo después.

Para solucionar el problema del tiempo de la animación, se decidió guardar los fotogramas y a partir de ellos componer un video para ser reproducido después. Al componer el video, se establece la velocidad de muestreo en fotogramas por segundo. Además, el programa también mide la duración de la animación en fps, por lo que se genera el número adecuado de fotogramas que corresponden a la duración de la animación. El video muestra una animación fluida y que además tiene la duración adecuada. Siguiendo el ejemplo de la palabra *hola*, el video correspondiente sí tiene una duración de un segundo que es lo que se esperaría.

Cabe señalar que el sistema se desarrolló y probó en una computadora portátil con las siguientes características: procesador AMD Athlon(tm) 64 X2 Dual-Core Processor TK-57, 1.90GHz, 2.00Gb de memoria RAM y sistema operativo Windows Vista de 32 bits. El programa ocupa cerca de 120Mb de memoria RAM, principalmente debido a que los 39 modelos 3D son cargados al iniciar la ejecución. En el apéndice A, se presentan los requerimientos mínimos del sistema y un manual de instalación. En teoría, el sistema debe funcionar sin problema en computadoras que cumplan estos requerimientos y los tiempos de la animación deben ser parecidos a los que se obtuvieron en las pruebas realizadas.

6.4. Uso del Sistema como Auxiliar en la Terapia del Lenguaje

El sistema se pensó desde un principio para ser usado en la terapia del lenguaje de niños con hipoacusia. Por ello, en verano de 2009 se presentó en la Escuela de Educación Especial Jean Piaget en la ciudad de Puebla. El sistema fue evaluado por la profesora María Isabel Haza Rubí, coordinadora de los grupos de audición de este instituto. La Profesora Haza Rubí consideró que el sistema puede utilizarse como herramienta auxiliar en la terapia de lenguaje, principalmente como herramienta para el maestro, ya que algunos de los niños aún no pueden escribir muy bien, o bien su vocabulario es muy reducido, lo cual impide que escriban en la interfaz del sistema. Por otro lado, las técnicas de animación utilizadas, en particular, la proyección de fonemas a visemas, va de acuerdo con las técnicas de lectura labio-facial que utilizan. Además, la Profesora Haza Rubí destacó que, gracias a que la interfaz permite rotar el rostro, se puede practicar la lectura de labios desde distintos ángulos, lo cual no permitían otros sistemas con los que han trabajado.

En general, el sistema fue considerado potencialmente útil para la terapia de niños con problemas de audición y se pretende trabajar con él en el próximo ciclo escolar. El programa se puede utilizar como parte de diversas dinámicas, como lo señaló la coordinadora. Por ejemplo, se puede jugar *Lotería* con los niños y en lugar de gritar las cartas que van saliendo, se puede usar el programa para que los niños a través de la lectura de labios, identifiquen la carta que ha salido y coloquen una ficha en su plantilla, si es que contiene esa carta.

La interfaz gráfica de usuario fue modificada para que fuera más fácil de utilizar por profesores y alumnos. En esta versión del sistema, se eliminaron las opciones de generar el video de la animación y de sincronización con audio, ya que requieren de configuraciones más avanzadas. Se pretende instalar el sistema en agosto de 2009, para que se empiece a utilizar en el laboratorio de cómputo del instituto por profesores y alumnos de los grupos de audición.

Conclusiones

En este proyecto de investigación, se desarrolló un sistema de animación facial utilizando la técnica de metamorfosis de modelos tridimensionales con interpolación lineal. A partir de un conjunto inicial de modelos 3D, se logró exitosamente generar nuevos modelos combinando los originales usando una metamorfosis ponderada. La animación se realiza interpolando en el tiempo, las posiciones de los vértices en el espacio tridimensional de un modelo fuente y un modelo destino. Se determinó utilizar la técnica de interpolación lineal por ser la más sencilla de implementar y tener un bajo costo computacional, además de que, la topología de los modelos 3D con los que se trabajó permite establecer una correspondencia uno-a-uno entre los vértices de las mallas poligonales y por lo tanto, no se presentan deformaciones indeseables al utilizar la interpolación lineal para las transiciones.

Los tiempos de la animación son asignados de dos maneras: por sílaba y de acuerdo a un archivo de audio. Para la asignación por sílaba, fue necesario diseñar e implementar un algoritmo de silabificación para el idioma español, en particular, el de México. El algoritmo, fue diseñado en base al estudio de fonología de este idioma realizado por Rafael A. Núñez Cedeño y Alfonso Morales-Front[51]. El algoritmo consta solamente de cuatro pasos: la identificación de núcleos silábicos, silabificación del núcleo con la consonante anterior, maximización de ataques silábicos y adjunción de codas; esta característica hace que su implementación sea sencilla y que el algoritmo en sí, constituya una aportación importante del trabajo. Cabe señalar, que el desempeño del algoritmo es bastante aceptable, comparado con otros algoritmos de silabificación del español[60][52].

Una parte importante del trabajo, es la proyección de fonemas del español a visemas. Existen ya, sistemas parecidos al desarrollado aquí pero que están hechos para idiomas diferentes al

español[35], o bien, son modificaciones de sistemas extranjeros[55]. Por ello, se considera que el presente trabajo contribuye de manera significativa al desarrollo del área de animación y gráficos por computadora en nuestro país, en particular, en el área de animación facial. La correspondencia entre fonemas del español y visemas que se propone aquí, arroja resultados satisfactorios en la animación del rostro. La profesora María Isabel Haza Rubí, coordinadora de los grupos de audición de la escuela de educación especial Jean Piaget, consideró que las técnicas de animación empleadas, son apropiadas y congruentes con las técnicas de lectura de labios que emplean en este instituto. Además, juzgó al sistema potencialmente útil como auxiliar en la terapia del lenguaje, utilizada en la rehabilitación de niños con hipoacusia.

Un objetivo adicional del proyecto, era sincronizar la animación con un archivo de audio correspondiente a la pronunciación del texto de entrada. Aunque se experimentó con un sintetizador de texto a voz para que el sistema generara también el audio correspondiente, los resultados no fueron del todo satisfactorios: la voz del sintetizador sonaba muy artificial y para sincronizar los tiempos del audio y la animación, se hubiera requerido estudiar a fondo la programación del sintetizador que se probó. Como alternativa, se utilizó la aplicación SAPI 5.1 Lipsync de la empresa Annosoft, para obtener los tiempos de los visemas a partir de un archivo de audio existente. Esta aplicación, escribe en un archivo de texto los tiempos de cada fonema, los cuales son leídos por el programa y son aplicados a los visemas correspondientes. Los resultados obtenidos son satisfactorios, ya que el video de la animación puede ser reproducido simultáneamente con el archivo de audio original y estar bastante bien sincronizados. Desgraciadamente, se tuvieron problemas al tratar de incorporar esta aplicación a nuestro sistema, por lo cual, para sincronizar la animación con una pista de audio es necesario realizar algunos pasos ajenos al sistema, como grabar la pista de audio, crear un archivo de texto con su transcripción y procesar estos datos con el programa SAPI 5.1 Lipsync para generar el archivo de tiempos que puede ser leído por nuestro programa.

También, para añadir expresividad al rostro, se implementó un sistema de modificación de expresiones faciales. Se reservaron secuencias de caracteres para modificar la expresión del rostro mientras pronuncia un determinado fragmento del texto. La interfaz gráfica de usuario contiene botones para ingresar iconos que representan a estas secuencias de caracteres reservadas, como se hace en los populares programas de mensajería instantánea, esto provee una interfaz familiar

y fácil de usar al usuario.

El usuario puede modificar la apariencia de la escena tridimensional, cambiando el color de ojos y piel del modelo, así como el color y textura de fondo de la escena. Con esto, se logra una mayor interacción entre el usuario y la animación, además de que al poder cambiar la apariencia como se desee, se elimina la monotonía del sistema.

Como conclusión final se puede decir que, el presente trabajo constituye una aportación significativa no solamente en el campo de la animación, sino también, en el campo de la educación especial, en particular, en la educación de niños con hipoacusia que practican la lectura de labios, además de que, la interacción entre humanos y computadoras puede ser mejorada por agentes animados que exhiben habilidades sociales-emotivas y/o asociaciones socio-culturales y que parecen tener vida. Para lograr estas metas el modelado y animación del rostro, expresiones faciales, voz, estilo visual y la personalidad resultante de tal agente juegan un papel vital y ofrecen una amplia oportunidad para una investigación multi-disciplinaria.

Limitaciones

Las limitaciones del sistema se pueden dividir en dos: las que tienen que ver con el funcionamiento interno y las técnicas de animación empleadas, y las limitaciones externas al sistema, como restricciones de hardware y software. En cuanto a las primeras, se puede decir que aunque el sistema cumple de manera aceptable los objetivos planteados, existen algunos detalles de implementación que presentan algunos problemas. Para empezar, se tiene una restricción en cuanto a los modelos tridimensionales del rostro que se utilizan, ya que éstos fueron conseguidos en la red y no fueron modelados específicamente para representar visemas en español y sobre todo, no fueron diseñados pensando en ser aplicados en la terapia del lenguaje. Por lo tanto, la animación del rostro está restringida a ser generada por transiciones entre estos modelos o sus combinaciones.

El proceso de análisis fonológico presenta limitaciones que son más difíciles de superar. Como ya se ha visto, diseñar un algoritmo de silabificación que silabifique correctamente todas las palabras es muy difícil, debido a las excepciones que existen a las reglas más generales. También, en la silabificación intervienen factores como las fronteras morfológicas de las palabras,

que para tomarse en cuenta, sería necesario contar con una base de datos de palabras, o realizar un tipo de análisis semántico del texto. Otros factores que intervienen en la silabificación y son ignorados por el sistema, son la coarticulación, la sinalefa y la sinéresis. La primera, ocurre cuando la pronunciación de un fonema es afectada por el contexto que lo rodea; las otras dos son procesos de resilabificación de grupos vocálicos, la sinalefa ocurre en las fronteras de palabras, mientras que la sinéresis reduce los hiatos en el interior de una palabra.

Con respecto a la proyección entre fonemas y visemas, se tiene la limitante de que si se desea construir un visema nuevo, éste debe poder ser representado como una combinación lineal de los modelos 3D originales. Esto afecta sobre todo, cuando el contraste visual del visema está dado principalmente por la lengua, ya que los modelos 3D con los que se trabajó tienen movimientos muy restringidos de la lengua.

Por otra parte, como se expuso en el Capítulo 6, los tiempos de la animación no coinciden con los esperados y es necesario componer un video para reproducirlo con los tiempos adecuados. Esto nos lleva también a pensar en las limitaciones externas al sistema, ya que una computadora con menos capacidad de memoria o una tarjeta de gráficos de menor calidad, puede resultar en una animación aun más lenta debido al tiempo de procesamiento.

Perspectivas

Como objetivo final del trabajo, se pretende aplicar el programa como auxiliar en la terapia del lenguaje, usada en la rehabilitación de niños con hipoacusia. El sistema será probado en la escuela de educación especial Jean Piaget de la ciudad de Puebla, donde ya ha sido presentado a la coordinadora de grupos de audición, la profesora María Isabel Haza Rubí, quien lo consideró apropiado y potencialmente útil para el instituto. Por ello, se atenderán algunas sugerencias de la profesora, en cuanto a las funciones del sistema, como por ejemplo, añadir a la interfaz la opción de ocultar el texto de entrada, para que los niños no lo vean y tengan que depender totalmente de la lectura de labios para entender el mensaje.

Entre las mejorías que se pueden hacer al sistema, se encuentra principalmente la sincronización con audio, ya que éste es un proceso que debe pasar por una etapa externa al sistema, a saber, la obtención de tiempos mediante al aplicación SAPI 5.1 Lipsync. Una al-

ternativa, es incluir de alguna forma esta etapa dentro del sistema, cuidando la sincronización entre los procesos. Otra opción, es volver a experimentar con un sintetizador de texto a voz, esta vez estudiando a fondo la programación del sintetizador y tal vez, mudando el sistema a un ambiente de desarrollo que sea más compatible tanto con el sistema como con el sintetizador.

Finalmente, se pretende buscar los espacios para presentar el trabajo en conferencias y publicarlo como artículo, ya que representa una contribución al campo de la animación facial y la educación especial en México.

Apéndice A

Manual de Instalación y de Usuario

En este apéndice se discuten los requerimientos de software y hardware del sistema, el proceso de instalación, y se presenta un breve manual de usuario.

A.1. Requerimientos del Sistema

El programa fue desarrollado usando la interfaz de programación de aplicaciones Java 3D, su versión 1.4.0_01 ha sido puesta en circulación para Solaris (sparc y x86), Linux (x86 y amd64), y Windows (32-bit).

Solaris/Sparc

La versión 1.4.0_01 de Java 3D para Solaris/SPARC requiere lo siguiente:

- JDK 1.4.2 o posterior (se recomienda 1.5.0) de Sun Microsystems: http://java.sun.com/j2se/
- Solaris 9 o posterior
- Frame Buffer con soporte para OpenGL (por ejemplo, XVR-1200)
- OpenGL 1.3 para Solaris o posterior. Para encontrar la versión actual, usar: `"pkginfo -l SUNWgrt"`. OpenGL para Solaris se puede obtener de:

  `http://www.sun.com/software/graphics/opengl/`

Solaris/x86

La versión 1.4.0_01 de Java 3D para Solaris/x86 requiere lo siguiente:

- JDK 1.5.0 o posterior de Sun Microsystems: http://java.sun.com/j2se/
- Solaris 10 o posterior
- NVIDIA Frame Buffer con OpenGL 1.3 o posterior

Linux

La versión 1.4.0_01 de Java 3D para Linux (x86 o amd64) requiere lo siguiente:

- JDK 1.4.2 o posterior (se recomienda 1.5.0) de Sun Microsystems: http://java.sun.com/j2se/
- Adaptador de gráficos con driver que soporte la extensión GLX: GLX 1.3 o posterior y OpenGL 1.2 o posterior.

Windows

La versión 1.4.0_01 de Java 3D para Windows 2000, y Windows/XP (32-bit) requiere lo siguiente:

- JDK 1.4.2 o posterior (se recomienda 1.5.0) de Sun Microsystems: http//java.sun.com/j2se/
- Windows 2000 o Windows/XP
- Soporte para OpenGL o DirectX como se muestra a continuación

Versión de OpenGL

El renderizador (renderer) OpenGL por default de Java 3D requiere OpenGL 1.2 o posterior, disponible del fabricante de la tarjeta de gráficos.

Versión de DirectX

El renderizador (opcional) DirectX de Java 3D requiere DirectX 9.0 o posterior, que está disponible por Microsoft en: http//www.microsoft.com/windows/directx/. La versión de DirectX de Java 3D es seleccionada al establecer la propiedad de sistema `"j3d.rend"` a `"d3d"`, por ejemplo:

```
java -Dj3d.rend=d3d ClassName
```

A.2. Instalación

Antes de instalar el programa, se debe instalar lo siguiente:

- Java SE Development Kit, JDK 1.5.0 (veáse la sección anterior sobre los requerimientos del sistema).
- Java 3D API: http://java.sun.com/javase/technologies/desktop/java3d/
- Java Media Framework API: http://java.sun.com/javase/technologies/desktop/media/jmf/

Una vez que se tiene instalado lo anterior, simplemente se corre el programa de instalación llamado AnimacionFacial, se siguen las intrucciones del instalador y al final el programa queda listo para usarse.

A.3. Manual de Usuario

Aquí se presenta un breve manual de usuario, la interfaz es fácil de utilizar por lo que no requiere de una explicación extensa. Al iniciar el programa se muestra la pantalla que aparece en la Figura A-1. A continuación se explican cada uno de los elementos de la interfaz.

Primero se tiene la barra de inicio de la aplicación (1), que contiene los botones para minimizar, maximizar y cerrar la ventana. El visor (2) muestra la escena 3D, al posicionar el ratón en esta ventana y arrastrarlo, se puede rotar el modelo 3D. También, al mover el scroll del ratón se puede acercar o alejar la cámara. El usuario debe escribir el texto a pronunciar en el panel de texto (3), los botones que se muestran en (4), sirven para añadir secuencias de caracteres reservadas al texto, para modificar la expresión del rostro. La primer ocurrencia de

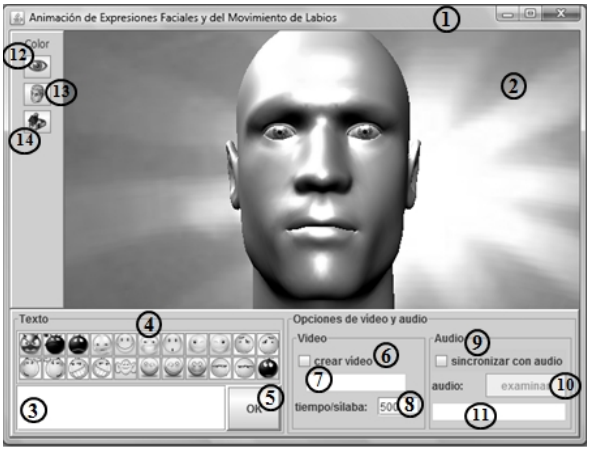

Figura A-1: Interfaz Gráfica de Usuario.

una de estas secuencias inicia el modificador de expresión en la animación, la segunda finaliza el modificador. Por ejemplo, si se tiene el texto *hola amigo*, pero se desea que la palabra *amigo* sea animada junto con una sonrisa, se debe escribir *hola =) amigo =)*, o bien, escribir *hola* luego presionar el botón que tiene una carita feliz, escribir *amigo* y volver a presionar el botón con la carita feliz. Las expresiones pueden estar anidadas, es decir, puede agregarse una detrás de otra. Debe recordarse que, un modificador de expresión estará activo hasta encontrar una segunda secuencia de caracteres de este modificador. Las secuencias de caracteres reservadas para los modificadores son las siguientes:

Enojo =[

Disgusto ={

Miedo =$

Tristeza =(

Alegría =}

Sonrisa =)

Sorpresa =*

Ojo izquierdo cerrado =‘

Ojo derecho cerrado =’

Ceja izquierda abajo =¿

Ceja derecha abajo =?

Ceja izquierda al centro =,

Ceja derecha al centro =.

Ceja izquierda arriba =¡

Ceja derecha arriba =!

Orejas hacia afuera ="

Aprieta ojo izquierdo =:

Aprieta ojo derecho =;

Ve hacia abajo =v

Ve hacia arriba =^

Ve hacia la izquierda =<

Ve hacia la derecha =>

No es necesario aprender de memoria estas secuencias, simplemente se pueden presionar los botones correspondientes en la interfaz para agregar los iconos al texto. La animación comienza cuando se presiona el botón que dice **OK** (5). Los tiempos de la animación se pueden asignar de dos maneras: por sílaba y por sincronización con un archivo de audio. Por default, los tiempos se asignan por sílaba, este tiempo se puede modificar al cambiar el valor de la caja de texto (8), el tiempo se debe escribir en milisegundos. Si se desea sincronizar la animación con un archivo de audio, se debe marcar la casilla (9), después de hacerlo se debe presionar el botón **Examinar** (10), y abrir un archivo de texto obtenido con la aplicación SAPI 5.1 Lipsync, la caja de texto (11) muestra el nombre del archivo abierto.

Si se desea grabar la animación en video, se debe marcar la casilla de verificación (6), y escribir en la caja de texto (7) un nombre para el video, sin extensiones ni rutas, la extensión del video siempre será .mov y se establece internamente. El video será guardado en la carpeta Videos que se encuentra en la carpeta raíz del programa.

Los botones (12), (13) y (14), sirven para cambiar el color de ojos, el color de piel, y el color de fondo respectivamente. Los dos primeros abren directamente un cuadro de colores para que el usuario seleccione uno, el botón (14) primero abre una ventana que permite al usuario decidir si desea asignar un color al fondo, o en su lugar usar una textura.

Apéndice B

Java 3D

En este apéndice se resumen los elementos principales de la interfaz de programación de aplicaciones (API) Java 3D. Esta API es usada para escribir applets y aplicaciones con gráficos en tercera dimensión. Provee una colección de constructores de alto nivel para crear, renderizar, y manipular una escena 3D compuesta de geometría, materiales, luces, sonidos, etc.

B.1. El Diagrama de la Escena (The Scene Graph)

Java 3D usa un diagrama de escena para organizar y manejar una aplicación 3D. La tubería de gráficos básica (*graphics pipeline*) está escondida, y es reemplazada por una estructura tipo árbol construida apartir de *nodos* que representan modelos 3D, luces, sonidos, el fondo, la cámara, y muchos otros elementos de la escena.

Los nodos tienen tipo, siendo la división principal entre nodos de grupo (**Group nodes**) y nodos hoja (**Leaf nodes**). Un nodo **Group** es uno que tiene nodos hijos, agrupándolos de manera que operaciones como traslaciones, rotaciones, y escalamientos puedan ser aplicados en masa. Los nodos **Leaf** son las hojas del diagrama, que a menudo representan los objetos visibles en la escena como los modelos, pero pueden ser entidades no tangibles, como luces y sonidos. Adicionalmente, un nodo **Leaf** puede tener componentes de nodo (*node components*), especificando el color, reflexión, y otros atributos de la hoja.

El diagrama de escena puede contener comportamientos (*behaviors*), siendo nodos que contienen código que puede afectar a otros nodos en el diagrama en tiempo de ejecución.

El término *diagrama de escena* o *gráfica de escena*, se utiliza en lugar del término *árbol de escena*, ya que es posible que los nodos se compartan, es decir, que tengan más de un padre.

La Java 3D API puede verse como un conjunto de clases que heredan de los nodos **Group** y **Leaf** en varias formas. La clase **Leaf** es subclasificada para definir diferentes tipos de figuras 3D y nodos de ambiente (*environmental nodes*), que representan luces, sonidos, y comportamientos. La clase de figura principal es llamada **Shape3D**, que usa dos componentes de nodo para definir su geometría y apariencia; estas clases son llamadas **Geometry** y **Appearance**.

La clase **Group** se encarga de la posición y orientación de sus hijos y se subclasifica para extender estas operaciones. Por ejemplo, **BranchGroup** permite añadir y remover hijos al diagrama en tiempo de ejecución; **transformGroup** permite cambiar la posición y la orientación de sus hijos [58].

B.2. !Hola Universo!

El ejemplo básico estándar para los programadores en Java 3D es **HelloUniverse** (presentado en el Capítulo 1 del tutorial de Java 3D de Sun). Muestra un cubo de colores rotando, como en la Figura B-1.

Figura B-1: Un cubo de colores que rota.

El diagrama de escena para esta aplicación se muestra en la Figura B-2. **VirtualUniverse** es el nodo raíz en todo diagrama de escena y representa el espacio del mundo virtual y su sistema de coordenadas. **Locale** actúa como la ubicación del diagrama de escena en el mundo virtual. Debajo del nodo **Locale** hay dos subdiagramas, la rama de la izquierda es la rama de

contenido (*content branch*), que tiene contenido específico del programa como geometría, luces, texturas, y el fondo. La rama de contenido difiere significativamente de una aplicación a otra.

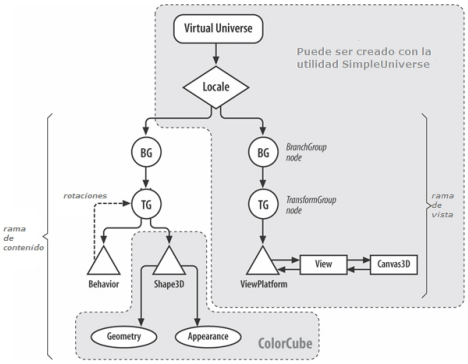

Figura B-2: Diagrama de escena del ejemplo HelloUniverse.

El cubo de colores, **ColorCube**, se compone de un nodo **Shape3D** y sus componentes asociados **Geometry** y **Appearance**. Su rotación se lleva a cabo por un nodo **Behavior**, que afecta al padre **transformGroup** de la figura **ColorCube**.

La rama de la derecha debajo de **Locale** es la rama de vista (*view branch*), que especifica la posición, orientación, y perspectiva del usuario para ver el mundo virtual desde el mundo físico. El nodo **ViewPlatform** guarda la posición del espectador (*viewer*) en el mundo virtual; el nodo **View** establece cómo transformar lo que el espectador ve, en una imagen del mundo físico (una imagen 2D en el monitor). El nodo **Canvas3D** es un componente de la interfaz gráfica de usuario (GUI) de Java que permite que la imagen 2D se coloque dentro de una aplicación de Java o un applet.

El **VirtualUniverse**, **Locale**, y la rama de vista frecuentemente tienen la misma estructura a través de distintas aplicaciones, debido a que la mayoría de los programas usan un único **Locale** y ven al mundo virtual como una imagen 2D en un monitor. Para estas aplicaciones, los nodos relevantes pueden ser creados con la utilidad **SimpleUniverse** de Java 3D, relevando al programador de mucho trabajo de construcción del diagrama [58].

Para más información de la interfaz de programación de aplicaciones Java 3D se pueden descargar los tutoriales de Sun Microsystems en:

`http://java.sun.com/developer/onlineTraining/java3d/`

Bibliografía

[1] RICK PARENT. Computer Animation: Algorithms and Techniques. Morgan Kaufmann Publishers. 2002. ISBN: 1-55860-579-7.

[2] NADIA MAGNENAT THALMAN y DANIEL THALMANN. Computer Animation. ACM Computing Surveys, Vol. 28, No. 1, Marzo 1996.

[3] MARCIA KUPERBERG et al. A Guide to Computer Animation for TV, Games, Multimedia and Web. Focal Press. 2002.

[4] P. BURNS. The Complete History of the Discovery of Cinematography. http://www.precinemahistory.net/introduction.htm, 2000.

[5] DAVID STURMAN. The State of Computer Animation. Computer Graphics. Febrero 1998.

[6] HACKATHORN, RONALD J. Anima II: a 3-D Color Animation System. Computer Graphics, proceedings of SIGGRAPH 77, ACM SIGGRAPH, New York, NY.

[7] REYNOLDS, CRAIG W. Computer Animation with Scripts and Actors. Computer Graphics, proceedings of SIGGRAPH 82, ACM SIGGRAPH, New York, NY.

[8] O'DONNELL, T.J y OLSON, ARTHUR J. "GRAMPS - A Graphics Language Interpreter for Real-Time, Interactive, Three-Dimensional Picture Editing and Animation". Computer Graphics, Vol. 15, No. 3. Agosto, 1981.

[9] G. STERN. Bbop: A Program for 3-Dimensional Animation. Nicograph 83, Diciembre, 1983.

[10] GIRARD, MICHAEL y ANTHONY A. MACIEJEWSKI. Computational Modeling for the Computer Animation of Legged Figures. Computer Graphics, proceedings of SIGGRAPH 85, ACM SIGGRAPH, New York, NY. 1985.

[11] WILHELMS, J. y B. A. BARSKY. Using Dynamic Analysis to Animate Articulated Bodies such as Humans and Robots. Graphics Interface 85 Proceedings, Montreal, Quebec, Canadá. Mayo, 1985.

[12] HAHN, JAMES K. Realistic Animation of Rigid Bodies. Computer Graphics, Vol. 22, No. 4. Agosto, 1988.

[13] BARAFF, DAVID. Analytical Methods for Dynamic Simulation of Non-penetrating Rigid Bodies. Computer Graphics, Vol. 23, No. 3. Julio, 1989.

[14] MCKENNA, MICHAEL y DAVID ZELTER. Dynamic Simulation of Autonomous Legged Locomotion. Computer Graphics, Vol. 24, No. 4. Agosto, 1990.

[15] REEVES, WILLIAM T. Particle Systems - A Technique for Modeling a Class of Fuzzy Objects. ACM Transactions on Graphics, Vol. 2, No. 2. Abril, 1983.

[16] BRUDERLIN, ARMIN y THOMAS W. CALVERT. Goal-Directed, Dynamic Animation of Human Walking. Computer Graphics, Vol. 23, No. 3. Julio, 1989.

[17] HODGINS, JESSICA K., WAYNE L. WOOTEN, DAVID C. BORGAN y JAMES F. O'BRIEN. Animating Human Athletics. SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer Graphics and interactive techniques. Septiembre, 1995.

[18] REYNOLDS, CRAIG W. Flocks, Herds, and Schools: A Distributed Behavioral Model. Computer Graphics, Vol. 21, No. 4. Julio, 1987.

[19] KASS, MICHAEL y GAVIN MILLER. Rapid, Stable Fluid Dynamics for Computer Graphics. Computer Graphics, Vol. 24, No.4. Agosto, 1990.

[20] PEACHEY, DARWYN R. Modeling Waves and Surf. SIGGRAPH '86: Proceedings of the 13th annual conference on Computer graphics and interactive techniques. Agosto, 1986.

[21] STAM, JOS y EUGENE FIUME. Turbulent Wind Fields for Gaseous Phenomena. SIGGRAPH '93: Proceedings of the 20th annual conference on Computer Graphics and interactive techniques. Septiembre, 1993.

[22] LI, XIN y J. MICHAEL MOSHELL. Modeling Soil: Realtime Dynamic Models for Soil Slippage and Manipulation. SIGGRAPH '93: Proceedings of the 20th annual conference on Computer Graphics and interactive techniques. Septiembre, 1993.

[23] TERZOPOULOS, DEMETRI, JOHN PLATT, ALAN BARR y KURT FLEISCHER. Elastically Deformable Models. Computer Graphics, Vol. 21, No. 4. Julio, 1987.

[24] MILLER, GAVIN S. P. The Motion Dynamics of Snakes and Worms. Computer Graphics, Vol. 22, No. 4. Agosto, 1988.

[25] GOLDENTHAL, RONY, DAVID HARMON, RAANAN FATTAL, MICHEL BERCOVIER y EITAN GRINSPUN. Efficient Simulation of Inextensible Cloth. ACM Transactions on Graphics, Vol. 26, No. 3, Artículo 49. Julio, 2007.

[26] LASSETER, JOHN. Principles of Traditional Animation Applied to 3D Computer Animation. Computer Graphics, Vol. 21, No. 4. Julio, 1987.

[27] GOVIL-PAI, SHALINI. Principles of Computer Graphics: Theory and Practice Using OpenGL and Maya. Springer, 2004.

[28] DUGELAY, JEAN-LUC, ATILLA BASKURT y MOHAMED DAOUDI. 3D Object Processing: Compression, Indexing and Watermarking. Wiley, 2008.

[29] CARR, J.C, et al. Reconstruction and Representation of 3D Objects with Radial Basis Functions. ACM SIGGRAPH 2001.

[30] FAIGIN, GARY. The Artist´s Complete Guide to Facial Expression. Watson-Guptill Publications. Nueva York, 1990. ISBN 0-8230-1628-5

[31] KÄHLER, KOLJA, JÖRG HABER y HANS-PETER SEIDEL. Reanimating the Dead: Reconstruction of Expressive Faces from Skull Data. SIGGRAPH '03. Julio, 2003.

[32] RADOVAN, MAURICIO y LAURETTE PRETORIUS. Facial Animation in a Nutshell: Past, Present and Future. Proceedings of SAICSIT 2006. Págs. 71-79.

[33] FORSEY, DAVID R. y RICHARD H. BARTELS. Hierarchical B-Spline Refinement. Computer Graphics, Vol. 22, No. 4. Agosto, 1988.

[34] DEROSE, TONY, MICHAEL KASS y TIEN TRUONG. Subdivision surfaces in character animation. SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques. Julio, 1998.

[35] EZZAT, TONY y TOMASO POGGIO. MikeTalk: A Talking Facial Display Based on Morphing Visemes.

[36] EKMAN P. y W. FRIESEN. Facial Action Coding System. Consulting Psychologists Press. Palo Alto, Calif. 1978.

[37] PARKE F. I. y WATERS K. Computer facial animation. AK Peters, Wesley, MA. ISBN 1-56881-014-8.

[38] SIFAKIS, E., NEVEROV, I, y FEDKIW, R. Automatic determination of facial muscle activations from sparse motion capture marker data. Proceedings of SIGGRAPH 2005, Skin & faces, 417-425.

[39] TERAN, J., SIFAKIS, E., SALINAS-BLEMKER, S., NG-THOW-HING, V. LAU, C., y FEDWIK, R. Creating and simulating skeletal muscle from the visible human data set. IEEE Transactions on Visualisation and Computer Graphics 11(3), 317-328.

[40] SURAZHSKY, VITALY y GOTSMAN CRAIG. Controllable Morphing of Compatible Planar Triangulations. ACM Transactions on Graphics, Vol. 20, No. 4. Octubre, 2001. Págs. 203-231.

[41] BEIER, THADDEUS y NEELY SHAWN. Feature-Based Image Metamorphosis. Computer Graphics, Vol. 26, No. 2. Julio, 1992.

[42] SEDERBERG, THOMAS W., PEISHENG GAO, GUOJIN WANG, y HONG MU- 2-D Shape Blending: An Intrinsic Solution to the Vertex Path Problem.

[43] COHEN-OR, DANIEL, DAVID LEVIN y AMIRA SOLOMOVICI. Three-Dimensional Distance Field Metamorphosis.

[44] ÁVILA JIMÉNEZ, MARÍA ELENA. Metamorfosis (Morphing). Tesis. Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación. Puebla, Pue. 2004. TES685.

[45] KENT, JAMES R., WAYNE E. CARLSON, RICHARD E. PARENT. Shape Transformation for Polyhedral Objects. Computer Graphics, Vol. 26, No. 2. Julio, 1992.

[46] LAZARUS, FRANCIS y ANNE VERROUST. Feature-Based Shape Transformation for Polyhedral Objects. Fifth Eurographics Workshop on Animation and Simulation. Oslo, septiembre 1994.

[47] GREGORY, ARTHUR D. Feature-based Surface Decomposition for Polyhedral Morphing. Proceedings of Computer Animation. 1998.

[48] HUGHES, JOHN F. Scheduled Fourier Volume Morphing. Computer Graphics, Vol. 26, No. 2. Julio, 1992.

[49] LERIOS, APOSTOLOS, CHASE D. GARFINKLE y MARC LEVOY. Feature-Based Volume Metamorphosis.

[50] FISHER, C. G. Confusions among visually perceived consonants. Jour. Speech and Hearing Research. 1968.

[51] CEDEÑO, NÚÑEZ RAFAEL A. y MORALES-FRONT ALFONSO. Fonología generativa contemporánea de la lengua española. Georgetown University Press. 1999. ISBN 0-87840-693-X

[52] DÍAZ ALONSO, JOSÉ LEOPOLDO. Prototipo de un Silabario Multimedia en el Idioma Español-Mexicano. Tesis. Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación. Puebla, Pue. Diciembre, 2004. TES718.

[53] INEGI. Las personas con discapacidad en México: una visión censal. ISBN 970-13-3590-2.

[54] ANDERSON, K. y MCOWAN, P.W. 2006. A real-time automated system for the recognition of human facial expressions. IEEE Transactions on Systems, Man, and Cyber-netics – Part B: Cybernetics 36(1), 96-105.

[55] RUIZ FLORES, LETICIA. Icatiani: Interfaz para Apoyo en Terapia de Lenguaje. Tesis. Universidad de las Américas Puebla. Otoño, 2000.

[56] LIMÓN ROSAS, MÓNICA. Intranet para Personas con Sordera. Tesis. Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación. Puebla, Pue. Marzo, 2003. TES624.

[57] SOMMERVILLE, IAN. Software Engineering. Addison-Wesley. ISBN 13: 978-0–321-31379-9.

[58] DAVISON, ANDREW. Killer Game Programming in Java. O'Reilly. Mayo, 2005. ISBN: 0-596-00730-2.

[59] ANNOSOFT PHONEME LABELS,

http://www.annosoft.com/sapi_lipsync/docs/group__anno40.html

[60] CUAYAHUITL, HERIBERTO, A Syllabification Algorithm for Spanish. Universidad Autónoma de Tlaxcala. CICLing 2004, LNCS 2945.