Paula Salgado Lucena

Expressive Talking Heads: Um Estudo de Fala e Expressão Facial em Personagens Virtuais

Dissertação apresentada ao Departamento de Informática da PUC-Rio como parte dos requisitos para a obtenção do título de Mestre em Ciências em Informática.
Orientador: Marcelo Gattass
Co-orientador: Luiz C. Velho

Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro

Rio de Janeiro, 25 de Junho de 2002.

Aos meus pais Lourdinha & Paulo,
a minha irmã Cynthia,
ao meu Rogério e a sua família.

Agradecimentos

A Deus pela vida e por ter me dado a capacidade de amar, aprender e sonhar.

Aos meus pais, Maria de Lourdes & Paulo, por ser fruto do amor deles, por tudo que sou e tudo que conquistei; a minha irmã, Cynthia, pelo exemplo de determinação, pela nossa união e amizade; e ao meu gato Grude, simplesmente por existir em minha vida.

A Rogério, namorado, amigo, conselheiro, orientador, professor; obrigada pelo seu amor, compreensão, paciência, cumplicidade, injeções de ânimo, nestes "quase 27 meses", enfim, por esta nossa conquista.

Ao Professor Luiz Velho, mestre e co-orientador da dissertação, por ter acreditado em mim e no meu sonho. Agradeço muito pelos ensinamentos passados não só na área acadêmica, mas nos ensinamentos de vida; pela paciência e persistência, pelos "puxões de orelha", por toda atenção, apoio, dedicação, entusiasmo, criatividade, inovação, infinitas idéias, enfim, por este trabalho e por muito do que aprendi até hoje.

Ao Professor Marcelo Gattass, mestre e orientador da dissertação, pelo apoio e ensinamentos ao longo deste mestrado, pelas oportunidades surgidas e por permitir e ajudar a realização deste meu sonho acadêmico.

Ao Pesquisador Ken Perlin, pela colaboração e orientação indireta do trabalho, e principalmente pela confiança depositada através do Prof. Luiz Velho no momento que investiu neste trabalho.

Ao Padre Josafa pelas missas às quintas-feiras. Por tantas vezes ter me passado ânimo e redobrado minhas forças, sem ao menos ter conhecimento de tudo que me passava.

À familia de Rogério, Mariana, Dona Vera & Sr. Gerson, avós, avôs, tios, tias, primos e primas, por todo apoio, carinho, torcida, simplesmente, por me fazer família.

A toda minha família, avós *(in memorian)* & avôs *(in memorian)*, tios, tias, primos e primas, pela torcida e apoio dados na distância física e no carinho sempre intenso quando estamos juntos.

Aos Professores do Mestrado, Bruno Feijó, Luiz Fernando Gomes Soares, Luiz Henrique Figueredo, Luiz Velho, Marcelo Gattass, Paulo Cesar Carvalho, Ralph Teixeira e Waldemar Celes, por todo conteúdo aprendido e que com certeza é de onde veio a base para este trabalho.

Aos Professores da Graduação (CIn-UFPE), em especial ao Professor Alejandro Frery, pelo incetivo para ir em busca deste sonho, pela torcida, pelo carinho e principalmente pela minha formação acadêmica.

Aos amigos da minha cidade Recife, Alexandre Coelho (Brabs), Conceição Lins (Ceiça), Danielle Valença (Dani), Eduardo Laureano (Du), Idevan Freire (Ide), Isabella Araújo (Bella), Janine Santos (Nine), Kátia Jordan (Ruiva), Larissa Mello (Lara), Maria Luíza (Malu), Patrícia Ramos (Pati), Sandra Valença (Sam), Sérgio Soares (Joca), ... (não iria parar de citar), pela torcida, pelas indas-e-vindas, pelo carinho, pela compreensão, simplesmente, por me completarem e serem peças fundamentais deste tesouro que é a amizade.

Às meninas que moram comigo, Luciana Lima (Lu), Lucimar Martins (Lu) e Viviane Braconi (Vi), pela amizade, por tudo que compartilhamos juntas, pelos ensinamentos de vida. Vivemos muita coisa em pouco tempo, mas apesar das turbulências, vencemos ("No final tudo deu certo...")!

Aos amigos da PUC, Adelailson Peixoto, Alesio Pfeifer, Alessandro Garcia, Antônio Carlos Azambuja (Caco), Aristófanes Correa, Diego Nehab, Elton Silva, Flávio Rodrigo, Gustavo Pierre e Taciana Melcop, pelos ensinamentos compartilhados, carinho, amizade e os bons momentos juntos.

Aos amigos do TecGraf, Camilo Fonseca, Eduardo Thadeu, Ivan Menezes e Sandra Schwabe por todo o apoio, torcida, sorrisos, fofocas, conselhos, enfim, pelos amigos que ganhei.

A Guilherme Cerqueira pelo constante suporte ao Linux, pela paciência e confiança.

A Antonio Scuri pelo profissionalismo, e por toda ajuda e ensinamentos passados.

A Luiz Cristóvão (Lula) e a Diego Nehab pelos conhecimentos da linguagem LaTeX repassados.

A Carlos Cassino pelo auxílio na linguagem Java e pelas oportunidades acadêmicas abertas, juntamente com Ana Lúcia Moura e Maria Júlia Lima.

À Yedda Campos, a Claudinei Gouveia e a Herivelton Bernardes por todas as ajudas, e principalmente pela amizade e carinho diários.

À Isabela Farah, Deborah Golçalves, Ruth Sousa, secretárias do Departamento de Informática/PUC-Rio, e a Nair Duarte, secretária do Visgraf/IMPA, pela atenção e carinho em todas as vezes que precisei.

À Carolina Alfaro pela paciência, disponibilidade e sabedoria desprendidas durante a revisão do texto desta dissertação.

Aos Professores Bruno Feijó e Hugo Fuks por terem avaliado este trabalho e por todas as críticas e sugestões que foram de fundamental importância.

A todos os demais amigos e companheiros do TecGraf/PUC-Rio, Visgraf/IMPA e do Departamento de Informática da PUC-Rio que contribuíram direta ou indiretamente para realização deste trabalho.

Ao CNPq e a Fundação Padre Leonel Franca pelo auxílio financeiro.

Resumo

A face humana é interessante e desafiadora acima de tudo pela sua familiaridade. Essencialmente, é a parte do corpo utilizada para reconhecer indivíduos. Assim como a face, a fala é um importante instrumento na forma de comunicação do ser humano. Através da fala é possível externar pensamentos e, muitas vezes, ela indica o estado de ânimo em que uma pessoa se encontra. Juntos, fala e face são os principais elementos de interatividade entre os seres humanos. Contudo, reproduzir com naturalidade e fidelidade as peculiaridades destes dois elementos no universo computacional não é uma tarefa simples, constituindo-se em tópicos de pesquisa em diversas áreas, em particular na animação facial.

Entre os diversos tipos de sistemas de animação facial, destacam-se como diretamente relacionados a este trabalho aqueles que envolvem a sincronização da fala de um personagem com a animação da sua face. Sistemas desse tipo são conhecidos como *talking head* ou *talking face*.

Para o desenvolvimento de um sistema *talking head*, é necessário identificar as possíveis abordagens para a modelagem dos dois elementos básicos: fala e face. Os modelos utilizados irão influenciar não apenas a maneira como a animação é conduzida, mas a própria forma de interatividade do sistema. Uma contribuição importante deste trabalho é o estudo das possíveis abordagens e a proposta de uma taxonomia para a classificação de sistemas *talking head*.

A partir da taxonomia proposta e fazendo uso de uma determinada abordagem para cada parâmetro analisado, foi desenvolvida uma aplicação que recebe como entrada um texto contendo a fala e anotações de expressividade, gênero e idioma, e gera como saída, em tempo real, a animação de um personagem virtual enunciando o texto de entrada com o áudio e os movimentos faciais sincronizados. O sistema desenvolvido, denominado "Expressive Talking Heads", explora a naturalidade da animação facial e ao mesmo tempo busca oferecer ao usuário uma interface com interatividade em tempo real. O "Expressive Talking Heads" pode ser executado tanto no modo isolado (*stand alone*) como acoplado a navegadores *web*, tendo sido projetado e desenvolvido com a preocupação de oferecer uma solução independente da plataforma e do sistema operacional utilizados.

Abstract

The human face is interesting and challenging mainly because of its familiarity. Essentially, it is the part of the human body that is used to recognize individuals. As well as the face, the speech is an important instrument for human communication, allowing the exteriorization of thoughts and the definition of emotions. Together, speech and face are the main elements of interactivity among human beings. However, the natural and faithful reproduction of the peculiarities of these elements in the computational universe is not a simple task, constituting topics of research in diverse areas, particularly in facial animation.

Among the diverse types of facial animation systems developed, those that involve the facial animation of the virtual character combined with speech synchronization are distinguished as directly related to this work. These kinds of systems are known as *talking head* or *talking face*.

For the development of a talking head system, it is necessary to identify the possible approaches for the speech and face modeling. The models used will influence not only the way that the animation is performed, but will also affect the system´s interactivity. An important contribution of the present master thesis is the study of several possible approaches for the main elements and the proposal of taxonomy for the classification of the talking head systems.

From the proposed taxonomy and making use of one approach for each analyzed parameter, an application was developed that receives as input a text composed by the character´s speech and genus, language and emotion parameters, and it generates as output, in real time, the animation of a virtual character uttering the input text with speech synchronization and expressiveness. The system developed, called "Expressive Talking Heads", explores the naturalness of facial animation and it seeks to offer the user a real-time interactivity interface. The "Expressive Talking Heads" system can run as a stand-alone application or connected to web browsers. It was designed and developed to provide a platform - and operating system-independent solution.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Funções em *Scheme* e de Pseudo-Códigos

Capítulo 1

Introdução

A animação facial de personagens tem despertado um grande interesse nos últimos anos. Esta não é uma linha de pesquisa recente; esforços nesta área e pesquisas relacionadas com a animação da face no computador existem há mais de 20 anos. Mas por que animar a face humana?

A face humana é interessante e desafiadora simplesmente pela sua familiaridade. Essencialmente, ela é a parte do corpo que é usada para reconhecer indivíduos: é possível reconhecer uma face entre um número grande de faces similares e ser capaz de detectar várias diferenças sutis através da expressão facial. Por essas e outras razões, a face humana tem sido um tema de ampla investigação na comunidade científica; em particular, a habilidade de modelar a face e as nuanças de uma expressão facial é um desafio existente na área de Computação Gráfica.

Assim como a face, a fala é um importante instrumento na forma de comunicação do ser humano. É através da fala que o ser humano externa seus pensamentos, e muitas vezes apenas com a fala é possível deduzir o estado de ânimo em que a pessoa se encontra.

Juntas, a fala e a face são os principais elementos de interatividade entre os seres humanos. É com base neles que a maioria das pessoas troca idéias e compartilha emoções. Contudo, reproduzir com naturalidade e fidelidade as peculiaridades da fala e da face no universo computacional não é uma tarefa simples, constituindo-se em tópicos de pesquisa em diversas áreas, em particular na animação facial.

Entre os diversos tipos de sistemas de animação facial, existe um de importante destaque e que está ligado a este trabalho: são os sistemas de animação facial que envolvem a sincronização da fala de um personagem com a animação da sua face, conhecidos como sistemas *talking head* ou *talking face*. Um sistema *talking head* é capaz de analisar uma fala de entrada e extrair informações necessárias para realizar uma animação facial enunciando a respectiva fala. A Figura 1.1 oferece uma visão geral do funcionamento de um sistema *talking head* genérico.

Combinar a fala e a face em um sistema de animação facial, buscando obter uma saída expressiva, implica em gerar na animação um sincronismo entre esses dois elementos. Obter essa sincronização é uma tarefa essencial pois, caso contrário, há uma perda de naturalidade e interatividade facilmente percebida no momento da animação.

De posse dos elementos fala e face, um componente bastante interessante que pode ser adicionado com a intenção de enriquecer um sistema *talking head* é a expressividade. Nos sistemas de animação facial, é a expressividade que permite identificar na face o estado de ânimo do personagem e ainda explorar transições entre expressões faciais.

A expressividade em um personagem virtual não atua apenas na face, mas influencia também a determinação dos outros elementos. No momento em que é definido um estado de ânimo para o personagem de uma animação, esta decisão reflete na fala, através de mudanças

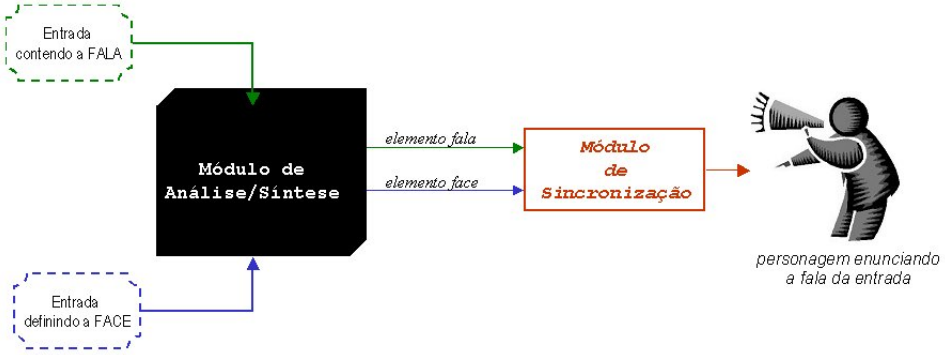


Figura 1.1: Visão geral de um sistema *talking head*.

na entonação e no ritmo da voz do personagem, e na face, através da posição e movimento dos componentes faciais para a composição da expressão desejada.

A expressividade é um componente que possibilita aumentar a naturalidade e o realismo oferecidos por um sistema *talking head*. No entanto, a utilização desse componente de forma adequada é uma tarefa difícil, de modo que o estudo e desenvolvimento de sistemas de animação facial com fala e expressões sincronizadas permanece como um tópico de pesquisa desafiador.

A construção de sistemas *talking head* possui basicamente dois aspectos que devem ser levados em consideração: a modelagem e a visualização. A modelagem consiste em utilizar dados abstratos empíricos ou aplicar técnicas de análise de dados reais para a construção do modelo facial, enquanto a visualização é definida a partir de uma síntese tendo como base o modelo construído. Existe uma série de possibilidades tanto para a modelagem quanto para a visualização dos sistemas *talking head* que serão discutidas ao longo deste trabalho.

É interessante observar que, para o caso particular dos sistemas *talking head*, não apenas os movimentos da boca e as expressões (visualização) podem ser derivados automaticamente de uma fala de entrada (modelo de entrada), como também a fala pode ser inferida a partir da visualização. Por exemplo, a partir da apresentação de um áudio e de operações de síntese, é possível deduzir como será a animação facial, ou ainda, a partir dos movimentos da face e de um mecanismo de reconhecimento é possível extrair o texto que está sendo pronunciado, através de algum esquema de leitura labial. No entanto, é importante salientar que, quando qualquer um dos dois componentes principais, fala e face, é retirado do processo, a qualidade da comunicação é degradada.

Um dos objetivos deste trabalho é pesquisar os elementos fala, face e animação, centrais em um sistema *talking head*. O estudo da fala refere-se aos seus fundamentos e às abordagens existentes para utilizá-la no mundo dos computadores, verificando que abordagem melhor se enquadra no contexto deste trabalho. A face é estuda de forma detalhada, visando conhecer seus principais elementos, suas etapas de definição e construção, e os aspectos relacionados à expressividade, descobrindo assim como ela possui vários níveis e formas de exploração. Finalmente, a partir da definição da face, as formas de animação são estudadas para melhor definir como dar vida a um personagem virtual.

Como mencionado acima, podem-se utilizar diferentes opções na definição dos elementos fala, face e animação, este último dependendo da forma com que a face foi definida. Portanto, outro tópico de pesquisa deste trabalho é o estudo das abordagens existentes para cada elemento principal. É estabelecido assim o segundo objetivo para este trabalho: propor uma classificação para os sistemas de animação facial com sincronização da fala. A taxonomia proposta constituirá uma das principais contribuições deste trabalho e abrangerá os parâmetros

(elementos) fala e face, além da forma de execução.

A partir da taxonomia proposta e fazendo uso de uma determinada abordagem para cada parâmetro analisado, este trabalho propõe o desafio de construir uma aplicação na qual um personagem virtual se comunicará com o usuário através de fala expressiva em um sistema multimídia. Sendo assim, o terceiro objetivo desta dissertação é descrever o desenvolvimento de uma aplicação que recebe como entrada um texto contendo a fala, anotações de expressividade, idioma e gênero, e gera como saída, em tempo real, uma animação desse personagem enunciando o texto de entrada com o áudio e os movimentos faciais sincronizados. Em decorrência dos objetivos aqui apresentados, o sistema desenvolvido foi batizado de "Expressive Talking Heads"[1].

Outro objetivo a ser contemplado pelo sistema "Expressive Talking Heads" é ser uma importante ferramenta de interatividade com o usuário. A fim de alcançar este propósito, o desenvolvimento do sistema busca caracterizá-lo como uma aplicação *web*. Para que o "Expressive Talking Heads" seja um sistema *talking head* voltado para a *web*, alguns requisitos adicionais precisam ser respeitados, como interatividade em tempo real e facilidade de instalação.

O restante deste capítulo destina-se a apresentar um pouco da história dos sistemas de animação facial, iniciando com a primeira "máquina falante", passando por sistemas simples e chegando até as grandes produções cinematográficas existentes nos dias de hoje. Adicionalmente, são abordados alguns dos campos comerciais e de pesquisa onde esses tipos de sistemas podem ser aplicados. Por fim, é apresentada a organização desta dissertação.

1.1 Histórico

O desejo de criar um sistema *talking head* de sucesso não é algo recente; pelo contrário, vem atravessando séculos. Atualmente, este desejo combina abordagens computacionais, cognitivas e biológicas, penetrando ainda em uma variedade de domínios e interesses.

A primeira "máquina falante" foi interpretada como um trabalho herético de magia. No século XIII, Albertus Magnus afirmou ter criado uma cabeça que poderia falar. Na verdade, o filósofo queria verificar se seria rotulado de herege por ter criado algo tido como abominável ou se sua reputação seria destruída por São Tomás de Aquino, um estudante conservador anterior a ele [RVB02].

Com o passar do tempo, enquanto algumas teorias sobre "máquinas falantes imitando o humano" eram derrubadas, outras cresciam e tomavam força. Nos dias atuais, essas "máquinas falantes" são poderosos sistemas de animação facial capazes de reproduzir, quase integralmente, a fala e as peculiaridades faciais dos seres humanos.

No que diz respeito aos trabalhos de representação facial baseados em computadores, o mais antigo data de 1972, quando foi criada a primeira animação facial tridimensional por Frederic Parke [Pea97].

A década de 1980 começou com o primeiro modelo facial baseado no controle dos músculos, sendo encerrada com novos modelos baseados em músculos e com abordagens para a sincronização automática da fala. Já os anos 1990 testemunharam o desenvolvimento das técnicas de animação facial por computador e sua utilização para grandes produções, como o filme de animação *Toy Story* [Lea95].

[1] "Expressive Talking Heads" faz referência ao sistema desenvolvido e "Expressive Talking Heads - Um Estudo de Fala e Expressão Facial em Personagens Virtuais" engloba todo o trabalho de pesquisa realizado, inclusive o sistema desenvolvido.

Se as tendências e evoluções passadas podem servir como indicadores para desenvolvimentos futuros, os próximos anos deverão ser bastante estimulantes para a área de animação facial por computador. Dirigido pelo crescente poder computacional, o desenvolvimento e o aprimoramento de técnicas para modelagem e animação de personagens representarão sempre desafios interessantes. Conseqüentemente, a quantidade e a qualidade da animação facial deverão crescer ao longo de vários segmentos. Mas onde aplicar estes sistemas?

As aplicações nesta área são inúmeras e de diversos escopos, indo desde ferramentas de pesquisa até trabalhos de criação de avatares para povoar espaços cibernéticos. Humanos virtuais e outras personalidades estão cada vez mais presentes em diversos lugares, incluindo filmes, televisão, brinquedos e jogos eletrônicos. Mas, além da pesquisa e da indústria de entretenimento, merecem destaque como campos de atuação dos sistemas *talking head* as áreas de fonoaudiologia e educação, por exemplo para ensinar a maneira de falar uma determinada palavra corretamente, e de vídeo-conferência, que visa fazer uma transmissão visual de um ponto a outro. Adicionalmente, estes sistemas atuam nas áreas de interface homem-computador e multimídia, em sistemas voltados para a *Internet* (por exemplo, ajudantes de navegação e vendedores virtuais), entre outros.

1.2 Organização da Dissertação

Este documento está estruturado conforme apresentado a seguir.

O Capítulo 2 destina-se a descrever os principais elementos que compõem um sistema *talking head*: fala, face e animação. Para cada elemento são mencionados os conceitos principais, além das técnicas existentes para a construção de um sistema *talking head*.

O Capítulo 3 propõe uma taxonomia para sistemas do tipo *talking head* através dos parâmetros fala, face e forma de execução. A partir das análises do capítulo anterior, é possível identificar as diferentes abordagens para os três parâmetros e com isto propor uma classificação para os sistemas *talking head*. Com a taxonomia definida, o capítulo apresenta os principais trabalhos em animação facial, relacionados ao "Expressive Talking Heads". Cada trabalho é classificado segundo a taxonomia proposta, possibilitando uma análise comparativa entre estes sistemas e o que o "Expressive Talking Heads" propõe desenvolver.

O Capítulo 4 tem seu enfoque na apresentação do sistema "Expressive Talking Heads" como um todo. O objetivo principal do sistema é explicado de forma mais detalhada e são apresentadas suas principais características. Por fim, é dada uma atenção especial aos subsistemas que compõem o "Expressive Talking Heads", sendo cada componente abordado separadamente.

O Capítulo 5 descreve os principais módulos de implementação do sistema, destacando a metodologia desenvolvida para integrar os diferentes subsistemas que compõem o "Expressive Talking Heads" e alguns dos algoritmos utilizados durante o seu desenvolvimento.

Por fim, o Capítulo 6 apresenta as conclusões deste trabalho, destacando as suas principais contribuições. Uma atenção especial é dada a trabalhos futuros que podem ser desenvolvidos tendo como base o "Expressive Talking Heads".

Capítulo 2

Elementos Principais de um Sistema *Talking Head*

Este capítulo apresenta uma análise dos principais elementos necessários para a construção de um sistema de animação facial com fala. O primeiro elemento analisado é a fala, que pode ser simplesmente vista como um sinal sonoro. Na Seção 2.1 são apresentados os fundamentos da fala que diretamente influenciam este trabalho e o método de geração (síntese) da fala.

Os outros dois elementos são a face e a animação. A relação entre eles é de dependência: a abordagem assumida para um indica que abordagem o outro deve assumir. A Seção 2.2 apresenta alguns dos conceitos básicos e características da face. Também é dada uma atenção especial à expressividade, um componente bastante enriquecedor para uma animação facial. Na Seção 2.3 são apresentados os principais conceitos da animação, desde a animação tradicional, passando pela animação por computador, até a animação facial. Por fim, são apresentados a relação e os requisitos necessários para a construção de um sistema *talking head* direcionado à *web*.

2.1 A Fala

O som é o fenômeno físico produzido pela vibração da matéria. À medida que a matéria vibra, as oscilações das partículas geram variações na pressão do ar ao redor [SN95]. Essa alteração na pressão faz com que as partículas adjacentes também oscilem, propagando o fenômeno adiante, formando um movimento de onda e gerando um sinal sonoro.

Sistemas multimídia tipicamente fazem uso de som apenas dentro do intervalo de freqüência do ouvido humano. Os sons dentro deste intervalo são chamados de *áudio* e as ondas neste intervalo são chamadas de *sinais acústicos*. Por exemplo, a fala é um sinal acústico.

A fala pode ser interpretada, entendida e gerada tanto por seres humanos como por máquinas. Uma pessoa se adapta de uma forma muito eficiente a diferentes oradores e a seus variados hábitos de fala. Apesar dos diferentes dialetos e pronúncias, a fala pode ser bem compreendida pelos humanos, pois o cérebro é capaz de distinguir o que é fala e o que é ruído [SN95].

Uma máquina pode suportar tanto a geração quanto o reconhecimento da fala. Na geração, a máquina possui a capacidade de sinteticamente gerar a fala, como o próprio termo indica. Os sinais gerados não possuem um som tão natural mas são facilmente entendidos. De forma contrária à síntese da fala, no reconhecimento são extraídas, a partir de um sinal acústico, as informações necessárias para a máquina. O conteúdo extraído possui uma voz mais natu-

ral comparada à voz humana, mas o seu entendimento pode ser mais complexo. Atualmente existem diferenças técnicas entre os sistemas de síntese (*speech synthesis*) [Dut97] e de reconhecimento (*speech recognition*) [Der98] da fala, as quais provavelmente perdurarão ainda por algum tempo.

A síntese da fala é o mecanismo de tratamento da fala utilizado no sistema "Expressive Talking Heads", portanto será abordada em maiores detalhes mais adiante. Já o mecanismo de reconhecimento da fala, apesar de bastante interessante, não se encontra no escopo deste trabalho. É apresentada apenas uma visão geral desta forma de tratamento e são mencionadas algumas das dificuldades inerentes hoje aos sistemas de reconhecimento da fala. A escolha do tratamento da fala através de sua geração em vez de seu reconhecimento está diretamente ligada ao objetivo deste trabalho, pois o usuário deve interagir com o sistema a partir de uma entrada textual. As características da síntese da fala contemplam de forma simples e direta este requisito do sistema "Expressive Talking Heads".

O reconhecimento automático da fala (ASR: *Automatic Speech Recognition* [Der98]) é um interessante tópico de pesquisa e possui uma extensa aplicabilidade. Sistemas com esse propósito são utilizados como ferramentas multimídia para suporte à navegação ou como mecanismos de inserção de dados. Embora as pesquisas nesta área datem de um longo tempo, os computadores ainda não são capazes de entender cada palavra que é pronunciada por qualquer pessoa. Portanto, os sistemas de reconhecimento de fala ainda são um desafio e um problema a ser solucionado.

Existem algumas dificuldades interessantes de serem salientadas. Uma das principais ocorre quando duas pessoas pronunciam a mesma palavra, pois cada falante tem sua forma particular de enunciar. Este problema é conhecido como "variação entre oradores" (*inter-speaker variation*). Uma segunda dificuldade é encontrada quando a mesma pessoa não pronuncia a mesma palavra identicamente em diferentes ocasiões, conhecida como "variação do orador" (*intra-speaker variation*) [Der98]. Ao contrário da máquina, o ser humano não vivencia esses dois tipos de problemas, pois o cérebro é capaz de reconhecer e tratar a ocorrência dessas situações.

O restante desta seção destina-se a apresentar alguns dos conceitos fundamentais da fala que serão de grande importância para o entendimento deste trabalho. Em particular, são abordados, de forma simples, os pontos cruciais do processo de síntese da fala, visto que este é o método de tratamento da fala utilizado no sistema "Expressive Talking Heads".

2.1.1 Fundamentos da Fala

A fala é um importante instrumento na forma de comunicação do ser humano, podendo ser naturalmente descrita através de propriedades fonéticas. Resumidamente, *fonemas* são os sons distintos em um idioma que o homem produz quando, pela voz, exprime seus pensamentos e emoções [Bec01]. Como exemplos de fonemas tem-se [*ka-za*] na palavra da língua portuguesa *casa* e [*w-uh-n*] na palavra da língua inglesa *one*[1].

A análise dos fonemas que compõem a fala tem sido uma estratégia comum nos sistemas de animação facial. Os fonemas são informações importantes, principalmente para o módulo de sincronização labial, pois permitem manter de uma forma mais precisa a sincronização da fala com a representação visual da face (*visemas*), como será abordado mais adiante neste trabalho.

Na realidade, com freqüência usam-se *difones* e *trifones* em vez de simples fonemas. Difones são pequenas seqüências de áudio amostradas através da transição do meio de um fonema para

[1]Os exemplos desta seção utilizam palavras da língua portuguesa e da língua inglesa, indistintamente.

o meio do fonema subseqüente. Pela estrutura fonética gerada, os difones foram escolhidos como a unidade de síntese para os sintetizadores concatenativos (Seção 2.1.2 e [Com00]). Existem entre 1.500 e 2.000 difones na língua inglesa, e o mapeamento do difone para o fonema é simples e direto. Por exemplo, a palavra inglesa *hello*, cuja representação em fonemas é */sil h ah l ou sil/*, pode ser mapeada na seguinte seqüência de difones: */sil-h/*, */h-ah/*, */ah-l/*, */l-ou/* e */ou-sil/* (/sil/ representa o fonema para o silêncio). Já trifones são coleções de três fonemas seqüenciais. Para um dado fonema, é considerado o fonema precedente e o fonema subseqüente. Na língua inglesa existem cerca de 10.000 unidades de trifones. Uma possível formação para trifones é a partir da concatenação de difones. Por exemplo, para os difones gerados na palavra *hello* no exemplo anterior, serão gerados os seguintes trifones: */sil-h-ah/*, */h-ah-l/*, */ah-l-ou/* e */l-ou-sil/*. A razão para utilizar difones e trifones advém da importância de capturar a dinâmica visual da fala, obrigando que os aspectos de coarticulação sejam considerados. Muitas vezes, a posição dos lábios para um mesmo fonema muda de acordo com o contexto em que o fonema se encontra inserido. Os efeitos de coarticulação estão presentes em diversas línguas.

É importante que se faça uma distinção entre o que é ouvido e como isto é representado ortograficamente. O fonema é uma realidade acústica, enquanto a *letra* é o sinal empregado para representar, na escrita, o sinal sonoro de uma língua. Por exemplo, a língua portuguesa possui em sua composição sete vogais orais tônicas, mas apenas cinco símbolos gráficos (letras): *a,e,i,o,u*. A Figura 2.1 ilustra o posicionamento dos lábios para algumas das vogais tônicas da língua portuguesa.

Figura 2.1: Posicionamento dos lábios para algumas das vogais tônicas da língua portuguesa.

Nesta dissertação, serão apresentados trabalhos que fazem uso de fonemas e trabalhos que fazem uso de difones e trifones para o tratamento da fala.

2.1.2 Síntese da Fala

Um sintetizador *Text-to-Speech* (TtS) é um sistema de computador capaz de ler em voz alta um texto fornecido como entrada [Dut97]. No contexto dos sistemas de síntese TtS é impossível gravar e armazenar todas as palavras de um idioma, de modo que é mais adequado definir um sintetizador TtS como um sistema de produção automática da fala através de uma transcrição dos grafemos para os fonemas correspondentes da sentença a ser pronunciada.

Todo sintetizador é resultado de uma imitação particular e original da capacidade humana de leitura, submetida a restrições tecnológicas e criativas que são características da época de desenvolvimento do sistema. Existe um grande número de aplicações para os sistemas TtS, merecendo destaque os serviços de telecomunicações, a área de educação e aprendizado de um

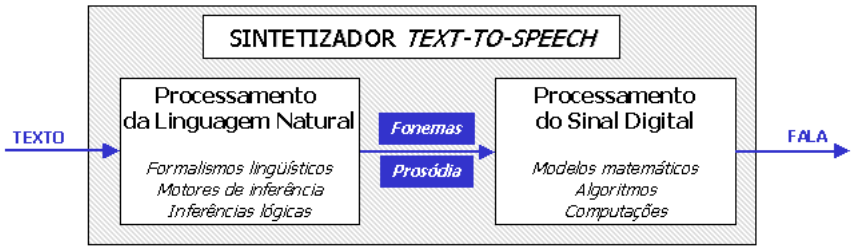


Figura 2.2: Visão geral do processo de síntese da fala.

idioma, a área de ajuda a deficientes físicos, o mercado de brinquedos e livros "falantes", a área de multimídia e, por fim, a área de pesquisa científica.

O processo de leitura para um computador procura ser o mais semelhante possível com o desempenhado por um ser humano, buscando a precisão e a naturalidade. Em um sistema TtS, o mecanismo de síntese da fala é realizado através de dois módulos, como ilustrado na Figura 2.2. O primeiro é um módulo de processamento da linguagem natural (NLP - *Natural Language Processing*), capaz de produzir uma transcrição fonética do texto lido, junto com a entonação e o ritmo desejados, isto é, a prosódia do texto[2]. O segundo é o módulo de processamento do sinal digital (DSP - *Digital Signal Processing*), que transforma a informação simbólica recebida do NLP em fala. Através dos módulos NLP e DSP, um sintetizador TtS tem a responsabilidade de gerar, a partir da entrada textual, um áudio como resultado da síntese. Outra saída útil para um sintetizador TtS, principalmente para os sistemas *talking head*, é a descrição dos fonemas.

Componente NLP

A Figura 2.3 ilustra uma estrutura proposta para um módulo de processamento da linguagem natural (NLP) genérico. O componente NLP é composto por um analisador textual que funciona em módulos: um pré-processador, um analisador morfológico, um analisador contextual e um *parser* sintático-prosódico. Além do analisador textual, o componente possui um módulo *letter-to-sound* e um módulo para a geração de prosódia. O restante desta seção destina-se a apresentar os módulos do componente NLP.

O módulo de pré-processamento é responsável pela organização das sentenças de entrada em uma lista de palavras manipuláveis. Ele identifica números, abreviações e acrônimos e os transforma em textos completos, quando necessário. O segundo módulo, o analisador morfológico, possui como tarefa propor todas as partes possíveis para as categorias de fala para cada palavra individualmente, tendo como base a ortografia. O analisador contextual, terceiro módulo do analisador textual, considera as palavras em seu contexto, reduzindo a lista das partes possíveis para as categorias de fala, gerada na etapa anterior, através da análise dos termos vizinhos. O último módulo do analisador textual é o *parser* sintático-prosódico, responsável por examinar o espaço de busca restante e encontrar a estrutura do texto, isto é, a organização em termos de cláusulas e sentenças, levando a uma aproximação com a prosódia esperada.

Uma vez finalizada a tarefa do analisador textual, o próximo módulo da unidade NLP é o *letter-to-sound*, responsável pela determinação automática da transcrição fonética do texto de entrada. Finalmente, há o módulo gerador da prosódia, no qual são levadas em consideração

[2]Entende-se por prosódia a parte da fonética que trata da acentuação e entonação corretas dos fonemas. A preocupação da prosódia é o conhecimento da sílaba predominante, conhecida como tônica [Bec01].

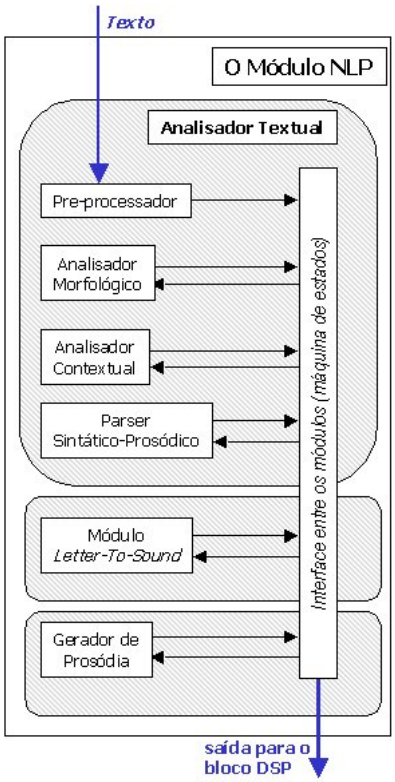


Figura 2.3: Visão geral do componente NLP.

propriedades que refletem no tom, na sonoridade e nas sílabas. Propriedades prosódicas possuem funções específicas na comunicação da fala, criando a segmentação da fala em grupos de sílabas e indicando a relação entre tais grupos.

Uma vez derivada a estrutura sintática-prosódica de uma sentença, ela é usada para obter a duração precisa de cada fonema (e dos silêncios), como também a entonação que deve ser aplicada.

Componente DSP

As operações do módulo de processamento do sinal digital (DSP) são executadas pelo computador de forma análoga ao controle dinâmico dos músculos articulares e da freqüência das cordas vocais pelos quais o corpo humano é responsável. Essas operações têm como objetivo gerar um sinal de saída que equivalha aos requisitos da entrada.

Visando funcionar de forma similar ao cérebro humano, o módulo DSP deve levar em consideração restrições do movimento de articulação. Assim, as transições entre os fonemas são bastante importantes para o entendimento da fala. Existem duas abordagens para o controle da transição fonética: a explícita e a implícita.

A transição fonética explícita ocorre na forma de uma série de regras que formalmente descrevem a influência dos fonemas, de um para os seus adjacentes. Já a transição fonética implícita ocorre através do armazenamento de exemplos de transições fonéticas e coarticulações na base de dados dos segmentos de fala, que são utilizadas simplesmente como unidades acústicas fundamentais (uso de difones e trifones, em vez de fonemas).

Essas duas abordagens deram origem a diferentes filosofias para a síntese. As divergências encontradas tanto no significado quanto nos objetivos levaram ao desenvolvimento da síntese por regras (*synthesis-by-rule*) e da síntese por concatenação (*synthesis-by-concatenation*).

Independente da forma de síntese assumida pelo componente DSP, por regras (explícita) ou por concatenação (implícita), o módulo DSP tem como objetivo final a geração do sinal de áudio referente ao texto fornecido como entrada para o módulo NLP. Em ambas as abordagens, a comunicação entre os módulos NLP e DSP é alcançada através de uma lista de segmentos (unidades) para fonemas.

A Figura 2.4 ilustra uma das abordagens para o processo de síntese: a síntese por regras. Esta figura fornece uma visão geral da síntese, tanto para o módulo NLP quanto para o DSP. O exemplo considerado é o de um sintetizador da língua portuguesa baseado em fonemas [SMA94], para a palavra *casa*.

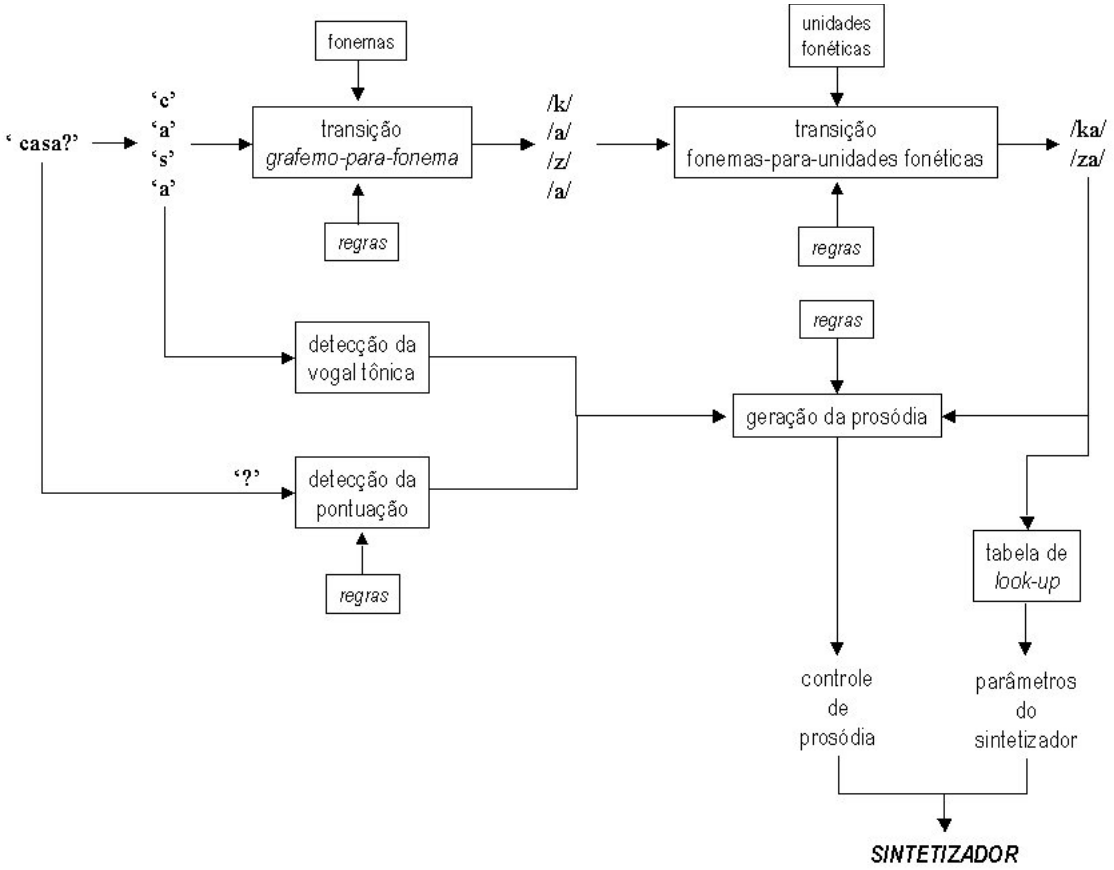


Figura 2.4: Exemplo de um esquema completo para a síntese.

2.2 A Face

A forma e as funções do corpo humano têm sido estudadas de modo detalhada por artistas e cientistas há séculos. Um exemplo particular é a era Renascentista, na qual a forma do corpo foi estudada, explorada e representada de diferentes maneiras. Atualmente, um dos objetivos de criar faces através do computador é não só buscar o realismo mas principalmente dar-lhes um movimento natural. Isto é contemplado através de seqüências animadas, diferente do que vem sendo feito desde o Renascimento, através de pinturas e imagens estáticas, ou mesmo nos primeiros desenhos animados, em que os personagens e seus movimentos eram limitados.

Esta seção destina-se a abordar os principais conceitos envolvidos na representação da face humana, com um foco particular no mundo dos computadores. Serão apresentadas, de forma

simplificada, algumas das características da anatomia facial humana que são pertinentes às questões de animação. Esta descrição é dividida em três partes: o esqueleto facial, os músculos faciais e a representação facial da fala. Por fim, um enfoque especial é dado aos elementos da expressividade facial.

2.2.1 Conceitos Básicos sobre a Face: Esqueleto e Músculos Faciais

A estrutura óssea da face humana é basicamente formada pelo crânio e pelo esqueleto facial. O crânio hospeda e protege o cérebro, enquanto o esqueleto facial é composto por um grande número de ossos, dos quais a mandíbula é o único que possui mobilidade. O esqueleto facial é a parte de interesse na modelagem da face para a grande maioria dos pesquisadores, pois provê o arcabouço onde os músculos e a pele estão localizados [PW96]. A Figura 2.5 (a) ilustra a estrutura óssea da face.

Os músculos são os órgãos dos movimentos. Através de contrações, eles movem várias partes do corpo, atuando inclusive na face. A energia de contração dos músculos é gerada mecanicamente por meio dos tendões, da aponeurose e da fáscia, que segura os terminais de cada músculo e controla a direção de suas trações. A relaxação dos músculos é passiva, ocorrendo através da carência de estímulos. Um dos interesses deste trabalho é pesquisar os músculos presentes na estrutura da face a fim de animá-la.

Os músculos da face são comumente conhecidos como *músculos da expressão facial*. Além de proporcionarem a própria expressão, alguns músculos faciais são responsáveis por outras funções, como, o movimento de bochechas e lábios durante a mastigação e a fala, a contração (constrição, fechamento) e dilatação (abertura) das pálpebras, entre outras. Os músculos da face trabalham de uma forma sinérgica, tornando-se difícil identificar e separar os limites de atuação de cada um. A Figura 2.5 (b) ilustra os músculos da face.

Como o número de músculos faciais e de detalhes é muito grande, este a estrutura facial desse trabalho ([Per97]) busca lidar com um número reduzido de músculos faciais que consigam produzir um conjunto de expressões faciais satisfatório, a fim de tornar a estrutura facial uma importante ferramenta de expressividade e de movimentos.

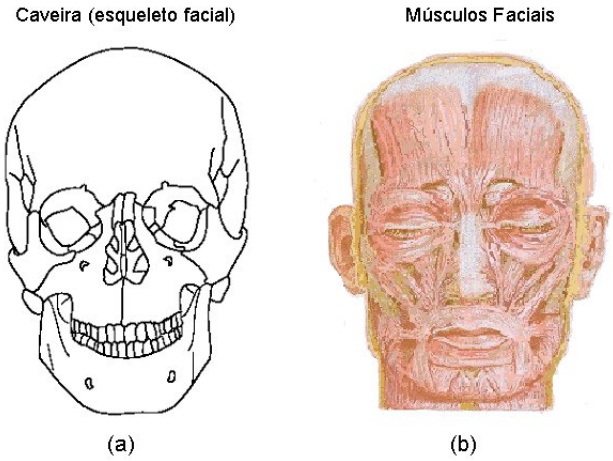


Figura 2.5: Esqueleto facial e músculos faciais.

Através dos músculos faciais é possível vincular a fala com a face. Cada fonema extraído da fala da pessoa pode ser representado visualmente através de uma expressão facial relativa a um certo formato da boca, conhecida como *visema*. De posse de uma base de visemas e da descrição fonética da fala, esta contendo não apenas a descrição dos fonemas mas também suas durações e

entonações, torna-se viável a construção das representações faciais e suas respectivas transições para cada pequeno trecho da fala. Cabe então a um outro módulo sincronizar a execução do áudio e a transição dos visemas, produzindo a animação facial enunciando o texto desejado. Esse módulo de sincronização é denominado *lip-sync* e será discutido posteriormente.

2.2.2 Modelagem da Face

O desenvolvimento de modelos faciais envolve determinar descrições geométricas que representem as faces de interesse, podendo ainda envolver atributos adicionais, como superfícies coloridas e texturas.

A face possui uma estrutura bastante complexa e flexível, apresentando ainda uma variação de texturas e cores, muitas vezes contendo inclusive rugas. É sabido que a anatomia detalhada da cabeça e da face é uma reunião complexa de ossos, cartilagens, músculos, nervos, vasos sangüíneos, glândulas, tecidos e pele, sendo que, neste trabalho, foi dada uma maior atenção ao conhecimento do esqueleto facial e ao entendimento dos músculos faciais. Até o momento não se tem conhecimento de algum sistema facial, construído a partir do computador, que faça uso do modelo facial completo. Felizmente, um grande número de aplicações, como a animação de personagens, podem ser efetuadas com modelos faciais que apenas se aproximam da anatomia facial. De fato, através de uma anatomia facial simples é possível construir modelos bastante ricos.

A modelagem de uma face e sua animação estão diretamente relacionadas. Escolhas feitas durante o processo de modelagem determinam a capacidade de animação e, reciprocamente, as ações que um modelo deve executar determinam como esse modelo deve ser construído. A etapa subseqüente à modelagem, a animação facial, será apresentada na Seção 2.3.

A mecânica da face e da cabeça é extremamente importante quando os modelos faciais estão sendo construídos. A mandíbula precisa trabalhar, as pálpebras devem se abrir e se fechar, além de outros pontos que precisam ser levados em consideração. Sem dúvida pode se afirmar que os olhos e a boca são as áreas mais expressivas da face. Através deles é possível comunicar um número grande de informações e extrair os melhores efeitos emocionais. Portanto, estas regiões devem ser tratadas no modelo facial com uma maior atenção.

Embora a maioria das faces possua uma estrutura familiar e um conjunto com as mesmas características, existem variações importantes de uma face para outra que são exatamente o que permite reconhecer as faces individualmente. Um dos desafios da animação facial é, justamente, o desenvolvimento de modelos que tratem essas variações.

A definição de uma geometria na modelagem da face objetiva representá-la de forma eficiente tanto no momento da visualização quanto no da sua animação. Algumas abordagens para a definição da geometria facial são aqui mencionadas.

Uma primeira possibilidade é utilizar uma das inúmeras técnicas de representação de volumes. Essa abordagem pode ser implícita, e inclui CSG (*constructive solid geometry*), vetores de elementos volumétricos (*voxel*) e elementos de volumes agregados (como *octrees*). Até o momento, esta não é uma abordagem popular para faces, sendo pouco explorada e aplicada, em decorrência da dificuldade de representação de faces realistas [PW96].

Uma segunda abordagem para a definição da geometria facial é através da representação de superfícies, sendo esta a representação geométrica atualmente preferida para modelos faciais. As estruturas da superfície normalmente devem permitir formatos de superfícies e mudanças nos formatos quando necessário, para várias faces e expressões conformes. Técnicas possíveis para a descrição de superfícies incluem superfícies paramétricas e superfícies poligonais, sendo esta última a técnica utilizada no sistema "Expressive Talking Heads". As su-

perfícies paramétricas incluem *B-splines* e *NURBS*, entre outras, enquanto as poligonais incluem malhas poligonais regulares e redes poligonais arbitrárias.

Representação Poligonal de uma Face

As estações de trabalho gráficas modernas são capazes de mostrar superfícies poligonais e atualizar modelos faciais de baixa complexidade em tempo real [PW96]. Devido a esta eficiência alcançada, a maioria dos modelos faciais é visualizada utilizando superfícies poligonais, como já foi mencionado. As superfícies poligonais podem ter a forma de malhas poligonais regulares ou de redes de polígonos arbitrários conectados, como ilustra a Figura 2.6. O uso de malhas poligonais regulares é incomum, devido à complexidade de construir uma face utilizando apenas polígonos regulares.

A topologia poligonal refere-se à forma como os polígonos são conectados para construir a superfície. A topologia de malha regular organiza os vértices dos polígonos em um vetor retangular. Esses vértices são então conectados com polígonos triangulares ou quadrilaterais para formar a superfície desejada. Já as redes de polígonos arbitrários são construídas através da conexão de vértices conforme o necessário para formar a superfície desejada.

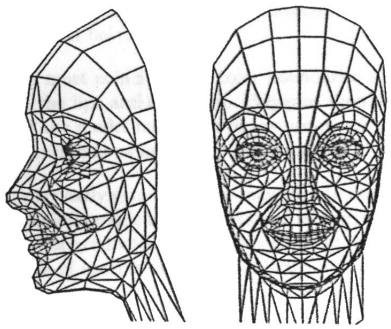


Figura 2.6: Exemplo de uma topologia poligonal (rede poligonal arbitrária).

Modelagem Baseada em Digitalização

Além da utilização das primitivas geométricas, existem outras técnicas que podem ser aplicadas para modelar uma face. Independente da abordagem utilizada, o objetivo da modelagem é a criação de descrições que representem fielmente a face desejada.

As faces também podem ser modeladas através de digitalizadores. Por exemplo, existem digitalizadores tridimensionais, que são dispositivos com um *hardware* especial capazes de localizar posições no espaço. Outro exemplo é o uso de *scanner 3D* para modelar a superfície poligonal de uma face.

Mas, de uma forma mais simples, as faces podem ainda ser definidas através de imagens capturadas que podem ser bidimensionais ou tridimensionais. A Seção 3.2 destina-se a apresentar comparativamente algumas das formas de definição de uma face (superfície poligonal e imagens capturadas).

2.2.3 Características dos Componentes Faciais

Um modelo facial é comumente formado pela soma de várias partes e detalhes. Muitas das pesquisas em animação facial têm focado em técnicas para definir e manipular geometricamente

a *máscara* facial. Uma abordagem mais completa consiste em integrar a máscara, os detalhes característicos da face e o cabelo em modelos que trabalham com toda a cabeça.

A máscara facial corresponde à parte externa visível da pele na superfície da face, não possuindo um realismo próprio. Normalmente é necessário adicionar outros elementos para torná-la convincente, como olhos, sobrancelhas, boca e orelhas, entre outros.

A dinâmica dos olhos é muito importante para obter sucesso na animação expressiva. Movimentos dos olhos e mudanças na dilatação das pupilas adicionam realismo à animação, enquanto detalhes de coloração da íris e reflexão das pálpebras dão mais vida à face.

A caracterização da boca se torna mais nítida se for feita de forma modular, através dos componentes lábios, dentes e língua. Os lábios e os tecidos ao redor são extremamente flexíveis e devem ser modelados com densidade de curvatura e flexibilidade suficientes para permitir todas as possíveis posturas da boca. A boca e os lábios contribuem decisivamente na comunicação de posturas emocionais. Os lábios são também um componente principal da fala, auxiliando na representação dos fonemas e, conseqüentemente, dos visemas. Os dentes são os componentes da boca que ficam parcialmente visíveis durante várias expressões e pronúncias da fala e, por sua vez, a língua fica visível em várias posturas faciais e é um elemento importante na distinção de posturas da voz.

Outro componente facial importante mas de difícil modelagem são as orelhas. As orelhas reais têm superfícies complexas que requerem um grande número de polígonos e superfícies bicúbicas com vários pontos de controle. Devido a esta complexidade, existem modelos faciais que não incluem as orelhas, "cobrindo-as" com cabelos, por exemplo.

Um último componente facial visto nesta seção é o cabelo. Compreendendo também sobrancelhas, cílios e barbas, ele é importante para modelos faciais realistas. O cabelo pode ser um aspecto importante da personalidade do personagem e tem sido um desafio para os animadores, devido à sua dinâmica e à representação diferente da superfície (as superfícies poligonais e cúbicas não são boas para representar o cabelo).

As características dos componentes faciais são inúmeras e de diferentes detalhes, e sua modelagem e animação enriquece a animação facial. Dentre os componentes faciais, dois, em especial, são explorados na modelagem e animação no "Expressive Talking Heads": a boca e os olhos[3]. A face deste trabalho (Seção 4.2.3 e [Per97]) busca através desses dois componentes manter os requisitos necessários para produzir uma animação satisfatória, sem causar perda na similaridade humana devido à ausência dos outros elementos.

2.2.4 Conhecendo a Expressividade

Para enriquecer o desenvolvimento de um sistema *talking head*, se faz necessário pesquisar os aspectos de expressividade da face. A expressividade é um elemento poderoso em uma face. Na animação facial por computador, é através dela que se torna possível identificar o estado de ânimo do personagem. Como o fator de expressividade é muito rico e possui muitas variações (a intensidade de cada expressão e suas possíveis combinações), este trabalho vai abordá-lo de forma preliminar, trabalhando apenas com as expressões faciais universais (simples), que serão apresentadas adiante.

As expressões faciais humanas têm sido o assunto de um grande número de investigações na comunidade científica. Em particular, a questão da universalização das expressões faciais através das diferentes culturas e as derivações de um pequeno número de expressões faciais

[3]O componente cabelo é modelado mas não possui animação, já os elementos orelhas e língua não são definidos.

principais têm consumido uma atenção considerável.

Pesquisas na área de expressão facial levaram à conclusão de que existem seis categorias universais de expressões faciais: tristeza, raiva, alegria, medo, desgosto e surpresa, como ilustra a Figura 2.7 [PW96]. Dentro de cada uma dessas categorias pode existir uma variedade de "intensidades" das expressões faciais e algumas variações nos seus detalhes. Há uma descrição proposta para cada uma das categorias de expressões faciais e suas variações em termos da aparência das três regiões faciais - as sobrancelhas, os olhos e a boca - e suas associações com as rugas faciais.

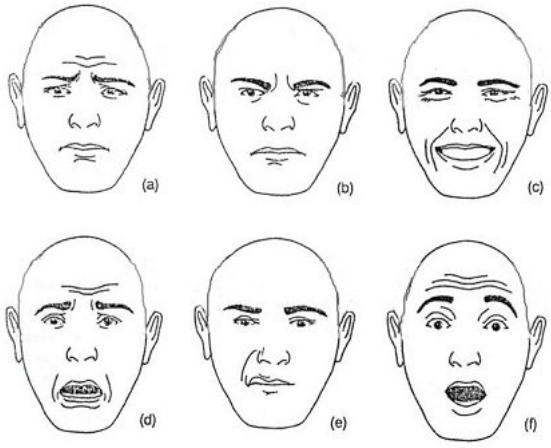


Figura 2.7: Expressões universais: (a) tristeza, (b) raiva, (c) alegria, (d) medo, (e) desgosto e (f) surpresa.

Cada uma dessas categorias possui traços característicos, principalmente nos componentes faciais olhos e boca, e nos locais onde as rugas se formam. Por exemplo, na categoria *alegria* (Figura 2.7 (c)) as sobrancelhas estão relaxadas. As pálpebras superiores estão levemente abaixadas e as inferiores estão retas, sendo puxadas para cima pelas bochechas superiores. A boca fica extensa, com os cantos puxados para trás em direção às orelhas. Se a boca está fechada, os lábios ficam finos; se está aberta, os lábios superiores ficam retos, mostrando os dentes superiores. Nesse caso, os lábios inferiores estão retos no meio e angulares próximos aos cantos. As rugas para a alegria são "pés de galinha" nos cantos dos olhos, uma ruga em forma de sorriso se fixa abaixo das pálpebras inferiores, aparecem covinhas na face e no queixo e uma ruga profunda nasolabial do nariz até o queixo. As variações para a alegria são gargalhada alta, gargalhada, sorriso com a boca aberta, sorriso, sorriso melancólico, sorriso ávido, sorriso ingrato, sorriso malicioso, sorriso debochado, sorriso com olhos fechados, sorriso falso e gargalhada falsa

Assim como a alegria, os traços característicos e as variações das demais categorias são descritos em [PW96].

A partir das categorias de emoções é possível dar vida aos personagens virtuais, enriquecendo muito os sistemas *talking head*. Existem ainda muitas expressões que não estão ligadas diretamente à emoção mas ao estado físico, tais como sofrimento (dor), sonolência e o ato de cantarolar.

Uma vez conhecidos os elementos fala, face e expressões faciais, o passo subseqüente para a construção de um sistema *talking head* é coordenar esses elementos, dando vida ao personagem virtual. Tal coordenação envolve a animação facial do personagem e a sincronização desta com a fala desejada.

2.3 A Animação

Completando a apresentação dos principais elementos para a construção de um sistema *talking head*, esta seção destina-se a abordar alguns dos conceitos da animação.

Primeiro será apresentada uma visão geral da evolução da animação, desde a animação tradicional, passando pela animação por computador, até a animação facial. Como o "Expressive Talking Heads" é um sistema de animação facial, uma atenção especial é dada a esta técnica. Esta seção termina enfocando a animação facial aplicada aos sistemas de animação facial de personagens virtuais com fala, apresentando alguns dos requisitos necessários para um sistema deste tipo voltado à *web*.

2.3.1 Animação Tradicional e Animação por Computador

A animação é o processo no qual os eventos e as características de uma cena mudam no decorrer do tempo. No caso particular da animação por computador, ela resume-se em utilizar uma ferramenta de desenho padrão para produzir quadros consecutivos, nos quais a animação consiste de movimentos relativos entre os corpos rígidos e movimentos possíveis que agem sobre estes corpos a partir de um ponto de vista ou de uma câmera virtual [WW92].

Historicamente, a animação vem sendo produzida de acordo com duas diferentes vertentes. A primeira é feita por artistas que criam uma sucessão de quadros de desenhos caricaturais, que são posteriormente combinados em um filme. A segunda consiste da utilização de um modelo físico - por exemplo, um boneco do *King Kong* - que é colocado em uma determinada postura e tem sua imagem capturada; em seguida é movimentado para uma nova posição e capturado novamente, e assim por diante, num processo contínuo para produzir a sensação de movimento.

As pesquisas realizadas nos anos 1980 na área de animação por computador mostram o desenvolvimento de técnicas bidimensionais e tridimensionais de animação baseadas na animação tradicional, também conhecida como animação convencional. Além de precursora, a animação convencional é a origem de vários princípios e termos utilizados na animação por computador [Las87].

De forma mais geral, quando se fala em técnicas de animação por computador, na verdade está se referindo aos sistemas de controle de movimento. Existem muitas complexidades e sutilezas envolvidas na especificação do movimento de objetos no espaço bidimensional e tridimensional. Ainda há diferentes caminhos através dos quais um problema pode ser abordado, não sendo possível afirmar que um dado sistema possui apenas um determinado tipo de controle de movimento durante sua animação ou se este controle é feito através de uma composição de técnicas.

2.3.2 Animação Facial

A animação facial teve início com os desenhos animados simples, alguns feitos à mão e outros tendo como base sistemas computacionais simples, nos quais os personagens não possuíam uma personalidade própria definida e normalmente só se trabalhava a silhueta. Freqüentemente, estes trabalhos eram desenvolvidos por animadores que tendiam a ser mais artistas do que cientistas da computação, de modo que a qualidade da animação produzida pelos sistemas computacionais da época dependia muito do animador-programador e da sua habilidade em construir a ponte entre a programação e o efeito visual desejado. Hoje em dia, existem sistemas bastante poderosos que permitem que o animador desenvolva as animações de forma cada vez

mais perfeita. É possível verificar esta evolução através dos famosos desenhos animados feitos em computador, como "Toy Story" [Lea95] e "Shrek" [Sea01].

O trabalho pioneiro em animação facial por computador foi desenvolvido por Frederic I. Parke nos anos 1970 [RVB02] [PW96]. O interesse neste tópico de pesquisa foi renovado no meio dos anos 1980 quando foi apresentada uma abordagem para a modelagem de músculos de expressões faciais por Keith Waters [PW96] [Pea97]. Subseqüentemente, vários sistemas foram surgindo, enriquecendo a área de pesquisa em animação facial por computador. Alguns destes trabalhos serão apresentados na Seção 3.4.

Como mencionado na Seção 2.2.2, a face humana pode ser definida através de imagens capturadas ou de um modelo geométrico. A animação da face através de imagens capturadas pode ser feita através da técnica de metamorfose de imagens (*morphing*), na qual, a partir de duas imagens I_1 e I_2, são geradas imagens intermediárias I_T, onde T é um parâmetro que varia entre 0 e 1. Essas imagens são geradas através da deformação (*warping*) de I_1 para I_2 e da deformação (*warping*) de I_2 para I_1, seguidas de um cruzamento dessas imagens (*cross-dissolving*).

Já para a face definida através de um modelo geométrico, existem ao menos cinco abordagens principais para sua animação: interpolação, baseada em performance, parametrização direta, baseada em pseudo-músculos e baseada em músculos [PW96]. O objetivo dessas várias técnicas é a manipulação da superfície da face em vários pontos, de forma a alcançar a expressão facial desejada em cada instante chave da seqüência de animação.

A técnica de interpolação é, provavelmente, a mais usada. Na sua forma mais simples, ela corresponde à abordagem *key-frame* (ou *key-pose*) encontrada na animação convencional. Nesta animação, a expressão facial desejada é especificada por um certo ponto (ponto chave) no tempo e então novamente para outro ponto no tempo alguns quadros depois. A partir dos pontos chave especificados, um algoritmo é capaz de gerar quadros entre esses quadros chave (*key frames*). Além de ser aplicada a modelos faciais, a interpolação é aplicada a superfícies flexíveis. A idéia por trás dessa técnica é que, dados dois valores, determina-se um valor intermediário especificado por um coeficiente de interpolação fracionária.

A técnica de animação baseada em performance supõe medições de ações reais do ser humano para dirigirem personagens sintéticos. Esta abordagem é normalmente utilizada em mecanismos de entrada interativos, tais como luvas de dados, roupas apropriadas e sistemas de *motion tracking* baseados em laser ou em vídeo.

Na técnica de parametrização direta, existe um desejo de criar um modelo encapsulado, o que gerará um vasto número de faces e expressões faciais baseadas num conjunto razoavelmente pequeno de parâmetros de controle de entrada. O objetivo é permitir que tanto as expressões faciais quanto as conformidades faciais sejam controladas por um conjunto de valores dos parâmetros. O ideal desta abordagem é ter um modelo que permita a definição de qualquer face possível com quaisquer expressões possíveis apenas através da especificação do conjunto de valores apropriado para os parâmetros.

Já na técnica baseada em pseudo-músculos, as ações dos músculos são simuladas usando operadores de deformação geométrica. Como foi mencionado na Seção 2.2, a interação complexa entre os tecidos, músculos e ossos resulta nas expressões faciais. É evidente que essas interações produzem um número enorme de combinações de movimentos. A idéia para a abordagem baseada em pseudo-músculos é desenvolver modelos faciais com poucos parâmetros de controle que emulam as ações básicas dos músculos da face, em vez de simular exatamente a anatomia detalhada da face. Esta técnica é a utilizada no sistema "Expressive Talking Heads", e os resultados de sua aplicação serão apresentados ao longo do Capítulo 5.

Por fim, na técnica baseada em músculos, é utilizado um modelo físico, podendo ser mola-

massa ou elementos finitos, para descrever os músculos faciais. Como a anatomia da cabeça e da face é bastante complexa, a idéia por trás desta abordagem é construir modelos que permitam a manipulação de expressões faciais baseadas na simulação de características dos músculos e tecidos faciais. Isto é possível através de um modelo no qual os vértices poligonais da superfície da face (a pele) são interconectados com modelos que possuem elasticidade. Esses vértices são então conectados à estrutura básica do osso usando músculos simulados. Esses músculos possuem propriedades elásticas, podendo gerar forças de contração. As expressões faciais são manipuladas através da aplicação de forças musculares à malha da pele elasticamente conectada.

É importante ressaltar que estas técnicas podem ser combinadas em um modelo geométrico da face. A Tabela 2.1 resume as principais abordagens apresentadas para a animação de uma face modelada através de uma malha poligonal.

Tabela 2.1: Abordagens para a animação facial para uma face definida através de uma malha poligonal.

Técnica	*Descrição Sumária*
Interpolação	a partir de dois valores determina-se um valor intermediário (base da animação *key frame*)
Baseada em Performance	medições de ações reais do humano dirigem o personagem sintético
Parametrização Direta	utilização de um conjunto de parâmetros para definir faces e expressões faciais
Baseada em Pseudo-Músculo	ações dos músculos são simuladas através de operadores de deformação geométrica
Baseada em Músculos	uso de um modelo físico para descrição dos músculos faciais

Independente da técnica utilizada na animação por computador, para construir uma animação facial de sucesso é importante conservar os princípios da animação tradicional [Las87] e aplicá-los na estrutura facial. Outro aspecto importante na definição da animação facial é a personalidade do personagem. Este não é propriamente um dos princípios fundamentais da animação tradicional mas consiste da aplicação inteligente de todos os princípios, sendo então um objetivo subjacente a todos eles.

Quando um personagem animado consegue prender a atenção da audiência, isso significa que o personagem e a história se tornaram mais importantes e aparentes do que a técnica por trás da animação, sendo um resultado bastante positivo. Na animação de um personagem, todas as suas ações e movimentos são resultados de processos estudados, senão seriam apenas uma série de movimentos sem relação. Através do personagem que pensa, é possível trazê-lo à vida.

2.3.3 Animação Facial e Sistemas *Talking Heads*

Como já visto, existem diferentes técnicas que podem ser aplicadas na construção de uma animação facial. Esta seção destina-se a apresentar uma das metodologias utilizadas no desenvolvimento de sistemas de animação facial com fala e componentes faciais sincronizados,

ou seja, os sistemas *talking head*. A metodologia aqui descrita é a que o "Expressive Talking Heads" utiliza no módulo de sincronização (Seção 5.3).

Além da sincronização para caracterizar um sistema de animação facial com fala, um dos requisitos que os sistemas *talking head* buscam durante a animação é a interatividade com o usuário. Existe ainda um subconjunto de sistemas *talking head* interativos que incorporam um requisito adicional, a execução em tempo real, visando disponibilizar a animação facial em tempo real na *World Wide Web*, mantendo a qualidade e a beleza da animação.

O "Expressive Talking Heads" é um sistema desenvolvido objetivando atender aos requisitos de interatividade e tempo real, projetado para ser executado tanto como uma aplicação local como uma aplicação *web* (Capítulo 5). Mas este é um subconjunto pequeno comparado ao conjunto dos *talking head*.

Existem alguns sistemas de animação facial com fala que não atendem a todos os requisitos mencionados (Seção 3.4). A maior parte dos sistemas *talking head* está preocupada apenas em contemplar o seu requisito básico: a sincronização da fala e da face. De uma forma geral, na construção de um sistema deste tipo a metodologia comum é aplicada no módulo de sincronização. Basicamente, a partir da apresentação do áudio (fala do personagem) verifica-se qual trecho de fala (fonemas) corresponde ao áudio no referido instante. Com esta informação é possível recuperar o visema apropriado e aplicá-lo na estrutura facial.

Com o mecanismo de sincronização definido, atender aos requisitos de interatividade, execução em tempo real e outros requisitos qualifica o sistema em questão como um *talking head* mais rico.

Requisitos para um Sistema *Talking Head* na *WWW*

A construção de um sistema *talking head* está intimamente ligada a alguns parâmetros, como a forma com que a fala do personagem virtual é reproduzida e a maneira como sua face é definida. Contudo, para a construção de um sistema *talking head* direcionado para a *web* há uma gama de requisitos que devem ser satisfeitos, e que serão aqui apresentados.

Tais requisitos são [Pan01]:

- Fácil instalação: em algumas aplicações, o personagem virtual não é a principal atração do serviço oferecido, então é necessário simplificar o procedimento de instalação.
- Qualidade visual: no caso de sistemas *talking head* virtuais na *web*, boa aparência e realismo são de extrema importância. No caso de se estar trabalhando com um personagem de estilo caricatural (*cartoon*), é possível ter maior liberdade no uso do lado artístico e no exagero da deformação, sabendo que a qualidade visual também deve ser boa.
- Rápido *download*: objetivando diminuir o atraso no *download* de um modelo, faz-se necessário desenvolver um modelo de baixa resolução e menor complexidade.
- Interatividade em tempo real: o personagem virtual deve ser capaz de interagir de diferentes formas dependendo do usuário e das ações que ele tomar. Neste caso, um *streaming* de vídeo e pré-processamento não são possíveis.

Capítulo 3

Uma Taxonomia para Sistemas *Talking Head*

No Capítulo 2 foram apresentados os principais elementos para a construção de um sistema *talking head*. Este capítulo destina-se a propor uma taxonomia para a classificação de sistemas *talking head*. Os três parâmetros escolhidos foram *fala*, *face* e *forma de execução*. O objetivo é utilizar esses três elementos para classificar as diferentes abordagens que podem ser utilizadas na implementação de sistemas *talking head*.

Adicionalmente, são apresentados alguns dos trabalhos pesquisados que se relacionam com a pesquisa desta dissertação. Os parâmetros e abordagens apresentados são utilizados na Seção 3.4 como ferramenta de comparação entre os trabalhos relacionados.

3.1 Fala

A fala em um sistema *talking head* está diretamente ligada ao áudio que é reproduzido junto com a animação facial, existindo duas abordagens que podem ser consideradas: o áudio da fala pode ser sintetizado ou capturado.

Na primeira abordagem, voz sintetizada, o áudio é gerado através de um sistema de *speech-synthesis* ou *text-to-speech*. Como já mencionado na Seção 2.1.2, um sintetizador *text-to-speech* (TtS) é um sistema de computador capaz de ler em voz alta um texto fornecido como entrada [Dut97], gerando a partir da entrada textual, o áudio digitalizado ou a estruturação fonética (possivelmente ambos).

A segunda abordagem faz uso da voz capturada. O procedimento usado na captura do áudio consiste em gravar a fala a ser reproduzida. O som capturado tanto pode provir de uma pessoa falando em um microfone como de um áudio já capturado e gravado que será reutilizado. Essa abordagem tende a proporcionar um efeito mais realista que a voz sintetizada, porém a voz capturada deve passar por um processo de reconhecimento (análise) para a extração da sua descrição fonética.

Independente da abordagem utilizada para a voz na animação, uma estratégia comum nos sistemas *talking head* tem sido a análise dos fonemas que compõem a fala. Os fonemas são informações importantes, principalmente para o módulo de sincronização labial, pois permitem que a sincronização da fala seja mantida de uma forma mais precisa com a representação visual da face (os visemas). Na realidade, o uso de difones e trifones é muito freqüente, em vez de simples fonemas, como visto na Seção 2.1.1. Devido a essa necessidade de análise dos fonemas, a fala sintetizada pode levar vantagem sobre a fala capturada, pois a análise dos

efeitos de co-articulação sobre os fonemas pode ser feita durante o processo de síntese. Já na fala capturada, essa análise pode ser feita apenas como uma etapa de pós-processamento, após a captura, utilizando mecanismos de reconhecimento de voz.

3.2 Face

No desenvolvimento de um sistema *talking head*, as etapas de modelagem e animação da face são complexas, mas de fundamental importância. Existem diversos fatores que devem ser considerados na construção e reprodução da face humana, entre eles a forma como a face é representada e o seu estilo, este implicando diretamente no grau de realismo aparente da face.

De maneira similar à fala, uma distinção possível para a representação da face humana é se ela é gerada a partir de uma imagem capturada ou se é definida através de um modelo geométrico (imagem sintetizada). Em ambas as abordagens, é possível definir uma face bidimensional ou tridimensional. Além disso, para os sistemas construídos a partir de um modelo geométrico, é possível fazer uso de texturas na face, produzindo um efeito de maior detalhe na imagem gerada.

No que se refere ao estilo que a face pode assumir em sua etapa de modelagem, pode-se ter um estilo realista ou um estilo caricatural, tanto para o modelo de imagens capturadas quanto para o modelo geométrico. O estilo realista busca ser o mais semelhante possível com a fisionomia da face humana, enquanto o estilo caricatural caracteriza-se pela presença de fatores de distorção ou exagero da face específica, sendo normalmente utilizado na produção de personagens em desenhos animados.

Uma vez modelada a face, um passo subseqüente é a sua animação. A animação, não apenas de faces mas de quaisquer objetos, está diretamente associada com movimento, ou seja, é um mecanismo que envolve dinâmica. No caso da face, essa dinâmica está vinculada a esforços aplicados sobre os músculos faciais, seja de contração ou relaxamento.

As diferentes abordagens de modelagem da face, com imagens capturas e modelo geométrico, possuem formas específicas para a animação da face depois de modelada. Na abordagem de face definida através de imagens capturadas, normalmente a produção da animação facial ocorre através da aplicação de técnicas de operação sobre imagens, mais comumente a técnica de metamorfose (*morphing*), explicada na Seção 2.3.2. Já para uma face definida através de um modelo geométrico, a animação é feita através da aplicação de técnicas de animação sobre os músculos faciais apresentadas na Seção 2.3.2. O objetivo dessas várias técnicas é a manipulação da superfície da face em vários pontos, de forma a alcançar a expressão facial desejada em cada instante chave da seqüência de animação. A malha poligonal, para ser animada, passa por operações de síntese de imagens, produzindo imagens que podem ser tanto no formato vetorial quanto no formato matricial.

O resultado final da animação de faces definidas através de imagens capturadas é um vídeo. Esta abordagem possui limitações com relação à expressividade, devido ao fato de que esta animação é feita através de metamorfose, sendo necessário aplicar a metamorfose a cada componente facial (olhos, sobrancelhas, bochechas etc.) de acordo com a expressão facial. Apesar da maior restrição em relação à expressividade, esta abordagem proporciona o realismo visual desejado, em decorrência de que as imagens capturadas representam exatamente o personagem virtual desejado.

A face definida através de um modelo geométrico possibilita um trabalho completo de expressividade, porque através dos vértices da malha é possível ter um maior controle sobre os músculos faciais. Para atingir uma maior semelhança visual com a face humana, é possível

aplicar texturas no modelo poligonal, como já mencionado.

3.3 Forma de Execução

Assim como as classificações apresentadas para a fala e a face, existem duas abordagens a serem consideradas no que diz respeito à forma de execução em um sistema *talking head*.

A primeira possibilidade consiste em executar a aplicação em tempo real. Em sistemas *talking head*, essa abordagem está diretamente ligada à interação com o usuário. A idéia por trás da execução em tempo real nesses sistemas é que, à medida que a entrada é fornecida pelo usuário (por exemplo, a digitação de um texto contendo a fala do personagem), é gerada a animação facial do personagem virtual enunciando o texto de entrada. É importante salientar que, nessa forma de execução, a inserção de dados deve poder ocorrer em paralelo à animação produzida como saída do sistema.

A segunda abordagem é a execução de um sistema *talking head* em *batch*. Ao contrário da abordagem anterior, este método não está vinculado à interatividade, caracterizando-se por ser uma abordagem passiva. A idéia central para sistemas *talking head* em *batch* é que uma dada entrada fornecida pelo usuário é posteriormente capturada e processada. A saída dessa etapa de processamento será a entrada para um módulo responsável por elaborar um vídeo da animação *talking head*, que será apresentado quando desejado.

3.4 Trabalhos Relacionados

Como mencionado na introdução deste documento, diversas pesquisas têm voltado sua atenção para a área de animação facial. Esta seção apresenta alguns dos trabalhos em animação facial relacionados ao sistema "Expressive Talking Heads" e encontrados na literatura.

De uma forma geral, em cada trabalho primeiro é apresentada uma descrição das principais características. Subseqüentemente, é feita uma classificação do trabalho segundo a taxonomia proposta neste capítulo, possibilitando assim uma análise comparativa entre eles e o que o "Expressive Talking Heads" propõe desenvolver.

3.4.1 Video Rewrite

O *Video Rewrite* [BCS97] é um sistema que faz uso de uma seqüência de vídeo existente para automaticamente criar um novo vídeo com o mesmo contexto mas com uma nova trilha sonora. Aplicações diretas de sistemas como o Video Rewrite são dublagem de filmes, teleconferência e efeitos especiais. Um exemplo recente do uso de técnica similar a essa é encontrado no filme *Forrest Gump*.

Basicamente, a criação de uma nova seqüência de vídeo é constituída de duas etapas: a análise para construção de uma base de dados de treinamento e a síntese da nova seqüência de vídeo.

A Figura 3.1 ilustra o estágio de análise do Video Rewrite. É responsabilidade desse estágio criar, a partir de uma seqüência original de um vídeo,um banco de dados de exemplos de quadros do vídeo que é também denominado de base de dados de treinamento. Na fase de análise, é possível que o sistema aprenda como a face de uma pessoa muda durante a sua fala. A dinâmica e a idiossincrasia da articulação da face, orientação da cabeça, formatos e posições da boca, dos maxilares e do queixo, entre outros, são então armazenadas na base de dados de

treinamento. Nessa etapa, são utilizadas técnicas de visão computacional para que o processo ocorra de forma automática.

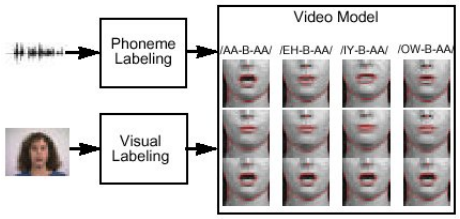


Figura 3.1: Visão geral do estágio de análise do Vídeo Rewrite.

Ainda na fase de análise, o Video Rewrite segmenta a trilha sonora original em fonemas e utiliza esses fonemas para rotular as imagens da base de dados de treinamento. Essa etapa de rotulação das imagens é a única do processo que necessita a intervenção humana. A coleção de exemplos de imagens rotuladas forma o que é então denominado de modelo de vídeo.

Na realidade, durante a fase de análise, o Video Rewrite segmenta a fala e o vídeo em trifones e não em simples fonemas, devido à necessidade de considerar os aspectos de coarticulação.

O estágio posterior ao de análise é o estágio de síntese, ilustrado na Figura 3.2. Nesse estágio, o novo áudio é segmentado e os fonemas obtidos são utilizados para selecionar a seqüência de vídeo contendo os trifones que mais se aproximam da nova trilha sonora. Tendo como base os rótulos definidos no estágio de análise, as novas imagens da boca são deformadas na face de fundo, sendo essa deformação feita através de técnicas de metamorfose (*morphing*).

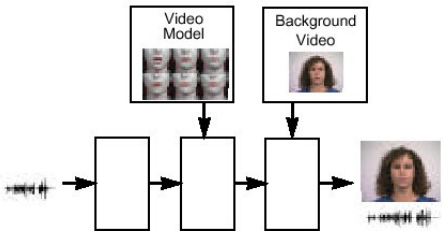


Figura 3.2: Visão geral do estágio de síntese do Vídeo Rewrite.

Como visto, o Video Rewrite é um sistema de animação facial que usa como elementos fundamentais um vídeo já existente e um novo áudio como entrada. A saída consiste da criação de uma nova seqüência realista do vídeo com o personagem enunciando a nova trilha sonora. O Video Rewrite combina basicamente a seqüência de vídeo de fundo, incluindo os movimentos naturais da face (como piscar dos olhos e os movimentos da cabeça) com novos movimentos para a boca e o queixo.

No que se refere à tanoxomia proposta, o Video Rewrite pode ser classificado como um trabalho em que a face é modelada através de imagens capturadas, o áudio é também capturado e não se trata de uma aplicação em tempo real. O Video Rewrite foi o primeiro sistema de animação facial a automatizar as tarefas de rotulação e montagem para resincronizar tomadas de vídeo existentes com novas trilhas sonoras. Esta automaticidade encontrada no Video Rewrite é uma importante referência para o sistema "Expressive Talking Heads" assim como para outros sistemas de animação facial.

3.4.2 MikeTalk

O segundo trabalho que pode ser mencionado é o *MikeTalk* [EP99] [EP98]. Este consiste de um sintetizador que converte texto em um fluxo de áudio, de forma similar a outros sintetizadores, mas que adicionalmente produz um fluxo visual enunciando o texto.

Esse trabalho é particularmente interessado em construir um sistema sintetizador de fala texto-visual (TTVS - *text-to-visual speech synthesis system*) onde a animação facial é vídeo-realista. A modelagem da face do personagem procura ser a mais parecida possível com a fisionomia humana, como se tivesse capturado a imagem através de uma câmera de vídeo, ao invés de lembrar um personagem caricatural. Outro aspecto do MikeTalk é que o trabalho concentra seus esforços no sistema visual do fluxo da fala e não na síntese do áudio. Para a tarefa de converter texto em áudio foi incorporado um sistema de síntese de fala denominado Festival [WT97]. De uma forma resumida, o Festival recebe como entrada um texto, que é sintetizado, gerando como saída um arquivo de áudio e um arquivo de fonemas, este último contendo informações de duração e entonação de cada fonema. No Capítulo 4 encontra-se uma descrição mais detalhada sobre o Festival e seu funcionamento, isto em conseqüência dele ser também uma das ferramentas TtS usadas no "Expressive Talking Heads".

Almejando atingir a peculiaridade de ser um sistema *talking head* vídeo-realista, o MikeTalk precisou que suas tarefas fossem cumpridas em duas partes. Um primeiro desafio foi o desenvolvimento de um módulo de fala visual (*Visual Speech Processing*). Este módulo tem como entrada os fluxos de duração e fonemas gerados pela unidade NLP (*Natural Language Processing*) do Festival e produz como saída um fluxo de fala visual enunciando o texto da entrada. O segundo desafio foi a construção de um módulo de *lip-sync* que sincroniza o áudio sintetizado com os fluxos visuais. A Figura 3.3 ilustra a visão geral do MikeTalk.

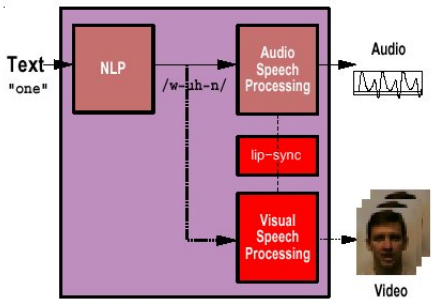


Figura 3.3: Visão geral do sistema TTVS MikeTalk

Como mencionado na Seção 2.2.4, um fator fundamental na construção de um sistema *talking head* é a natureza do modelo facial a ser usado. Dentre as abordagens existentes para modelagem de face, o MikeTalk baseia-se em imagens capturadas e faz uso de técnicas de metamorfose para a transição suave dos visemas.

Para a etapa de transcrição de fonemas para visemas, o MikeTalk assumiu inicialmente um mapeamento um-para-um, onde uma única imagem de visema é associada a cada fonema. Assim foi construída a base de extração dos visemas, chamada de corpo visual e ilustrada na Figura 3.4. Desta forma, existe ao menos uma palavra que uma vez pronunciada instancia cada um dos seus fonemas. Esta estratégia é bastante razoável e levou à extração de 52 imagens de visemas: 24 representando consoantes, 12 representando monotongos e 16 representando ditongos.

Em decorrência da similaridade de um grande número de visemas foi decidido aplicar uma redução no número dessas imagens agrupando-as através da comparação e semelhança.

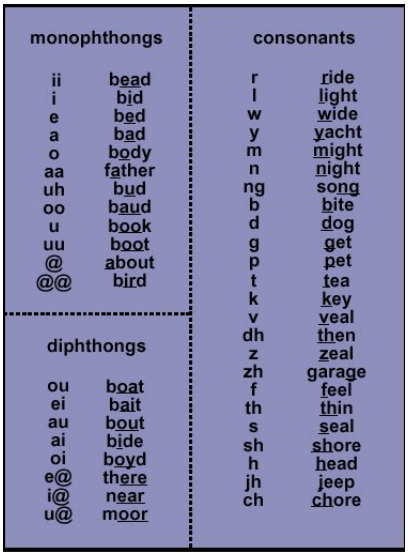


Figura 3.4: O corpo visual armazenado.

Em conseqüência foi constatado que o mapeamento de fonema para visema pode ser do tipo muitos-para-um: existem muitos fonemas que são parecidos visualmente, pertencendo assim à mesma categoria visual. No entanto, estudos também comprovam que este mapeamento pode ser considerado como um-para-muitos: o mesmo fonema pode apresentar diferentes formas visuais. Este fenômeno é conhecido como coarticulação. No MikeTalk, o efeito da coarticulação de fonemas não foi levado em consideração.

O conjunto final, após a redução de visemas, é mostrado na Figura 3.5. Este conjunto é formado por 16 visemas: 6 visemas representando os 24 fonemas consonantais, 7 visemas representando os 12 fonemas monotongos e 2 visemas representando os ditongos. Por fim, um último visema foi incluído para representar o silêncio.

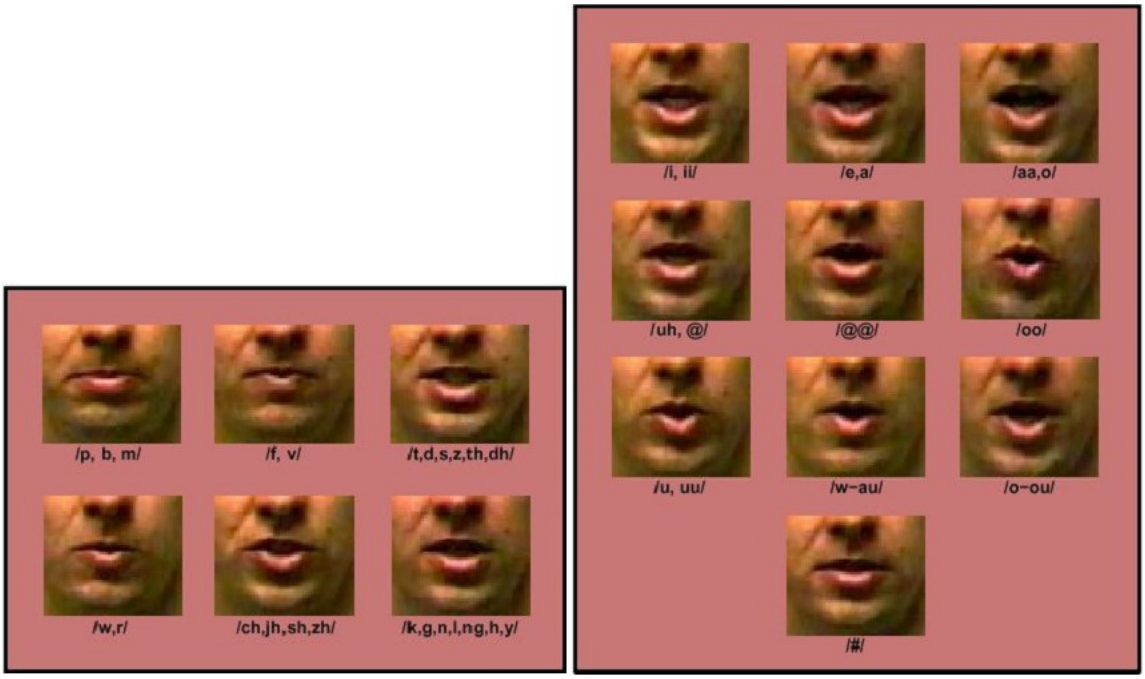


Figura 3.5: Os 6 visemas consonantais, os 7 visemas monotongos, os 2 visemas ditongos e o visema do silêncio.

Para se ter um sistema *talking head* de qualidade é extremamente importante que o mecanismo de transição de um visema para o seu subseqüente seja suave e realista. Esta é a etapa subseqüente à transcrição fonema-visema e é realizada através de técnicas de concatenação via metamorfose(*morphing*).

Após a concatenação dos visemas, a próxima etapa consiste em trabalhar o áudio. O Festival executa essa tarefa construindo o fluxo do áudio pela concatenação de difones. Cabe ao módulo de sincronização labial fazer com que a animação do personagem e o áudio sejam reproduzidos de forma sincronizada.

O módulo de sincronização labial cria um fluxo intermediário de transição de visemas. Assim, ele carrega a transição apropriada de visemas examinando qual difone do áudio está sendo apresentado em um dado instante de tempo. A Figura 3.6 ilustra o diagrama de sincronização labial.

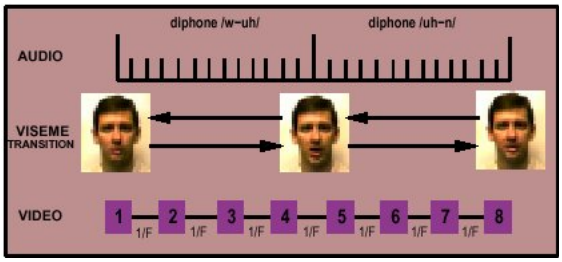


Figura 3.6: Diagrama de sincronização labial.

Dentre os trabalhos apresentados nesta dissertação, o MikeTalk é o sistema que mais se assemelha ao "Expressive Talking Heads". Em termos da taxonomia, o MikeTalk caracteriza-se como um *talking head* onde a imagem é capturada e vídeo-realista, a fala é sintetizada e sua execução não ocorre em tempo real.

3.4.3 Facade

O *Facade: Starford Facial Animation System* [DiP01] é um sistema de animação facial parametrizado que faz uso de uma topologia fixa de face poligonal. Esse sistema pode ser utilizado para criar e animar tipos faciais como também expressões faciais. A Figura 3.7 ilustra uma visão geral do sistema e seus vários módulos (ferramentas).

O Facade basicamente é composto por sete ferramentas: visualização, facial, animação e *rendering*, câmera, sincronização labial, texturização e iluminação. A ferramenta de visualização permite que a face do personagem seja exibida no modo aramado (*wireframe*), em um modo *flat shading* ou em um modo *smooth shading*.

A ferramenta facial no Facade é composta por um grupo de 51 parâmetros que uma vez modificados refletem sobre os olhos, a boca e a face. Através dessa ferramenta, o Facade pode ser usado para definir estados de ânimo que um personagem pode assumir ou para criar tipos de personagens, desde personagens de estilo realista buscando semelhança com o humano a personagens de estilo *cartoon* fazendo uso de parâmetros de exagero. A Figura 3.8 (a) ilustra um exemplo de emoção definida a partir da mudança de parâmetros para o personagem a partir do seu estado de neutralidade, enquanto que a Figura 3.8 (b) ilustra um personagem caricatural criado a partir da variação dos valores dos parâmetros faciais.

A ferramenta de animação e *rendering* possibilita a definição e o controle dos quadros chaves da animação (*keyframes*) e a apresentação da animação em uma seqüência de 30 quadros por segundo. A ferramenta de câmera permite a manipulação da câmera em diversas posições.

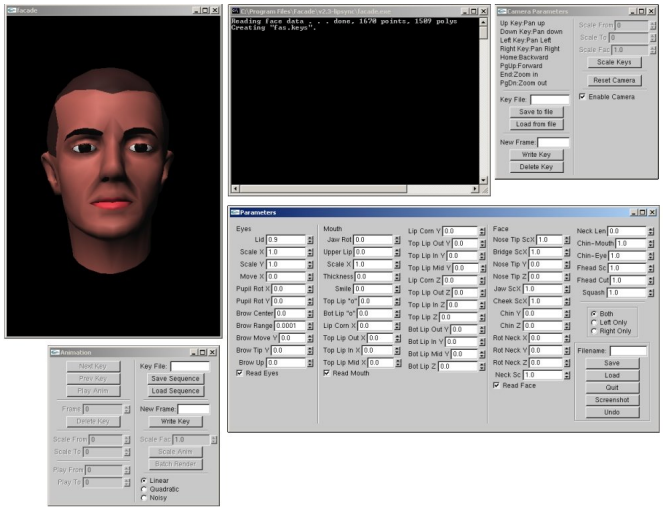


Figura 3.7: Visão geral do sistema Facade.

Figura 3.8: Expressões faciais e novos personagens definidos através de mudanças de parâmetros na ferramenta facial.

Para efetuar a sincronização labial da fala com a animação facial do personagem foi usada a ferramenta Magpie [GG00] para transcrição fonética manual que faz a análise da voz capturada. Esta ferramenta foi associada com uma ferramenta para a fala, o BaldiSync, que é um módulo do CSLU Toolkit [Hos92] para desenvolvimento de aplicações e pesquisas sobre a fala, responsável pela sincronização da fala arbitrária com movimentos labiais animados. A Figura 3.9 ilustra a ferramenta BaldiSync, usada para produção da fala. Uma vez construída a fala do personagem, a sincronização no Facade é feita de forma automática.

No Facade é possível adicionar textura na face do personagem, sendo que esta ferramenta de texturização ainda é bastante limitada, estando restrita ao formato *bitmap*. A ferramenta de iluminação permite que a face seja iluminada em planos diversos de uma forma bastante simples.

De acordo com a taxonomia para sistemas *talking head* proposta, o Facade caracteriza-se por ser um sistema que faz uso de um modelo geométrico para representar a face, a fala baseia-se em áudio capturado e, assim como os outros até então mencionados, sua forma de execução é em *batch*.

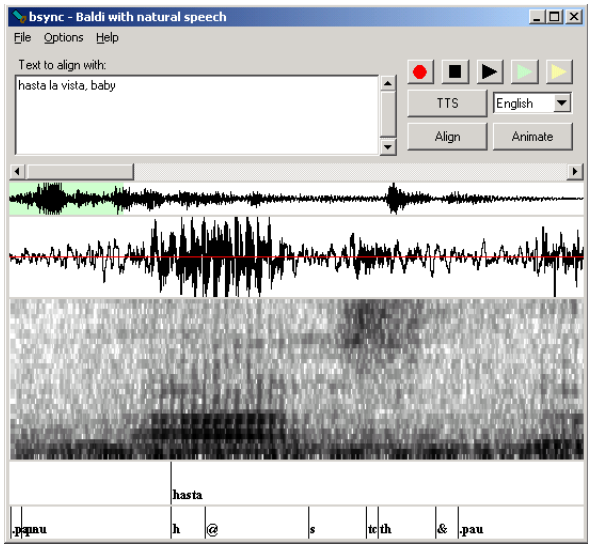


Figura 3.9: Ferramenta de sincronização labial utilizada no Facade.

3.4.4 FaceWorks

O *DIGITAL FaceWorks Animation Software* [Fac98] é um sistema de animação facial desenvolvido para aplicações multimídia [Fac98]. A animação no sistema ocorre através de uma face sintética 3D, cujos movimentos e expressões dos lábios são sincronizados com a fala real.

A face no FaceWorks é definida através de um modelo poligonal tridimensional onde são estabelecidos componentes faciais chaves para animação. A Figura 3.10 ilustra o editor geométrico no FaceWorks. Através dessa figura é possível ter ainda uma visão geral do sistema, onde a malha poligonal utilizada para definição da face é ilustrada mais à direita. À esquerda é possível manipular painéis para adição de textura sobre olhos, dentes, entre outras coisas. Texturas podem ser aplicadas ao modelo geométrico da face no FaceWorks objetivando alcançar um maior nível de detalhamento.

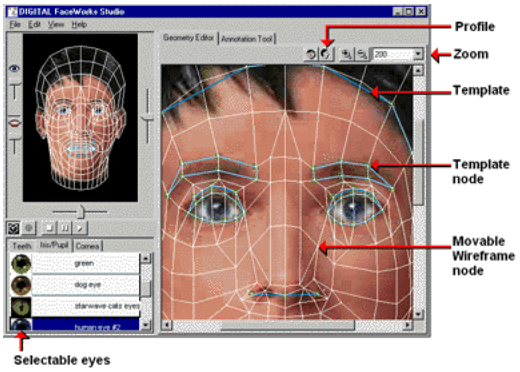


Figura 3.10: Visão geral do sistema DIGITAL FaceWorks e módulo de edição geométrica.

A animação facial no FaceWorks é feita sobre os músculos da estrutura poligonal da face. Um fator de relaxamento ou contração é aplicado aos músculos necessários para a animação desejada.

A voz do personagem é obtida através de um áudio capturado. O FaceWorks possui uma ferramenta para captura do áudio e é através deste módulo do sistema que é possível gravar o áudio e reproduzi-lo junto com a animação do personagem. Para ser capaz de sincronizar os movimentos dos lábios com o áudio capturado é necessário interpretar o arquivo de áudio para

obter os fonemas. A Figura 3.11 ilustra o editor de anotações do sistema. Através deste editor é possível acompanhar o processo de sincronização da fala e da animação, como ainda selecionar uma determinada região de reprodução da fala. Como ilustrado nessa figura, é possível verificar os fonemas gerados e a onda acústica. O FaceWorks permite também a seleção de um estado de ânimo para o personagem, a partir de um grupo de opções, como ilustra a Figura 3.11 em sua posição mais à esquerda e abaixo.

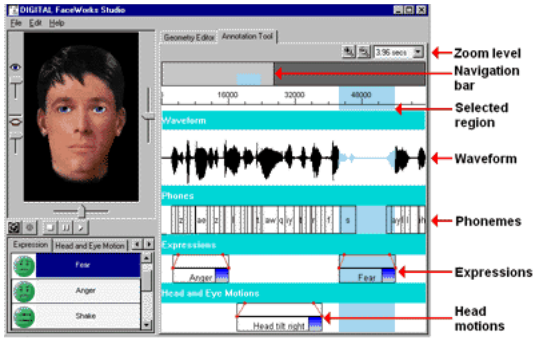


Figura 3.11: Editor de anotações no FaceWorks.

Uma vez gerada a voz do personagem virtual, através da captura do áudio, o FaceWorks possui um mecanismo de reprodução da animação facial do personagem estando a fala e as expressões faciais sincronizados. A Figura 3.12 ilustra a área de exibição da animação facial. A interface do sistema permite ao usuário observar as edições à medida que as mesmas ocorrem. Por exemplo, se alguns pontos chaves são manipulados na malha poligonal da face através editor geométrico, o *display* mostra o resultado desta modificação automaticamente. Isto também ocorre no editor de anotações, quando uma seleção do áudio é reproduzida a face fala automaticamente.

Figura 3.12: *Display* para reprodução da animação facial com fala e elementos faciais sincronizados.

Com a voz do personagem capturada, o FaceWorks simplifica o desenvolvimento da animação facial através da geração automática da sincronização do áudio, movimentos dos lábios, expressões e movimentos da cabeça.

De acordo com a taxononomia para sistemas *talking head* apresentada nesse capítulo, o FaceWorks é um *talking head* em que sua face é representada através de um modelo poligonal podendo fazer uso de textura e sua fala provém de um áudio capturado. Apesar do sistema refletir as modificações feitas pelo usuário em tempo real, não é possível qualificar o sistema como sendo de execução em tempo real. Isto devido ao fato de que o áudio é previamente

capturado. A partir do áudio é gerada a animação para posterior apresentação da mesma. A execução em tempo real está associada a interação do usuário com o sistema, de forma que o sistema interpreta e atua no exato instante da interação. Sendo assim, o FaceWorks é um sistema executa em *batch*.

3.4.5 Animação Facial baseada no Padrão MPEG-4

O trabalho *Designing MPEG-4 Facial Animation Tables for Web Applications* [GMT01] tem como objetivo apresentar uma técnica para auxiliar na construção de sistemas de animação facial na *web* baseados no padrão MPEG-4 [Koe01]. Na realidade, este não é um trabalho diretamente relacionado ao "Expressive Talking Heads", uma vez que não descreve um sistema *talking head* propriamente dito. O trabalho define um módulo de animação facial que pode ser utilizado na construção de outros sistemas. Um exemplo de aplicação proposto pelos autores para o módulo desenvolvido é a implementação de um sistema que possua uma face sintética, uma saída de áudio e uma entrada textual para permitir o diálogo com o usuário, ou seja, um sistema *talking head*. A Figura 3.13 ilustra a visão geral do exemplo sistema de animação facial proposto em referência [GMT01].

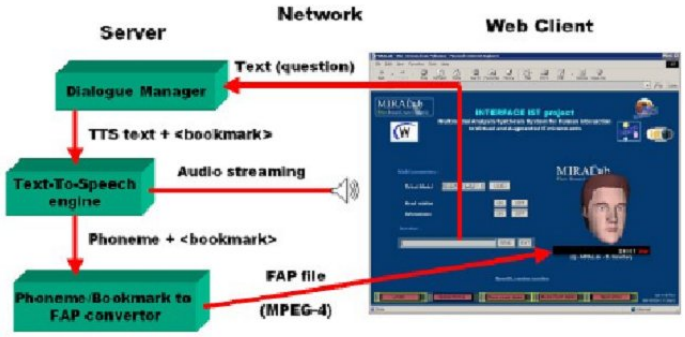


Figura 3.13: Visão geral do exemplo de aplicação proposto.

De fato, o que o trabalho *Designing MPEG-4 Facial Animation Tables for Web Applications* propõe como aplicação é exatamente o que foi desenvolvido neste trabalho. A diferença entre os dois sistemas é que, enquanto aquele propõe utilizar o padrão MPEG-4 para a modelagem e animação dos componentes faciais, este baseia-se em uma linguagem de marcação (Seção 5.1.2) para especificação da fala e das expressões faciais.

Para que o módulo de animação facial apresentado no *Designing MPEG-4 Facial Animation Tables for Web Applications* possa ser acoplado ao "Expressive Talking Heads", seria necessário que a especificação da animação facial fosse feita usando o padrão MPEG-4. No entanto, uma vez que este trabalho é resultado de uma cooperação com a Universidade de Nova York, optou-se por utilizar o subsistema *Responsive Face* (Seção 4.2.3 e [Per97]) como módulo para a modelagem e animação da face. Além de ser possível investigar a integração do módulo MPEG-4 desenvolvido em [Koe01] ao "Expressive Talking Heads", outro trabalho futuro seria implementar um motor MPEG-4 que permitisse utilizar o *Responsive Face* (e o próprio "Expressive Talking Heads") para realizar animações faciais especificadas em MPEG-4.

O padrão MPEG-4 foi formulado pela ISO/IEC JTC1/SC29/WG11 (MPEG - *Moving Pictures Expert Group*). Um subgrupo do MPEG-4, o SNHC (*Synthetic Natural Hybrid Coding*), idealizou um método de codificação eficiente para modelos gráficos e a transmissão comprimida dos parâmetros de animação específicos para o tipo do modelo. Para faces sintéticas, os

parâmetros de animação da face (FAP - *Facial Animation Parameters*) foram projetados para codificar animações de faces reproduzindo expressões, emoções e a pronúncia da fala. São definidos 68 parâmetros, categorizados em 10 diferentes grupos relacionados com as partes da face.

O padrão trabalha com visemas e expressões faciais, em um conjunto de ações faciais básicas. Foram definidos 14 visemas e 6 expressões faciais, ambos claramente distintos. As expressões faciais são baseadas nas categorias das expressões universais ilustradas na Figura 3.14. Com o objetivo de tratar os problemas de coarticulação da fala e do movimento da boca, transições de um visema para o seu subseqüente são definidas pela combinação dos dois visemas fazendo uso de um fator de contribuição (*weighting factor*) de cada visema. Ao contrário dos visemas, as expressões faciais são animadas através da definição de um fator de excitação mas, de forma análoga, duas expressões faciais podem ser combinadas através de um fator de contribuição (*weighting factor*).

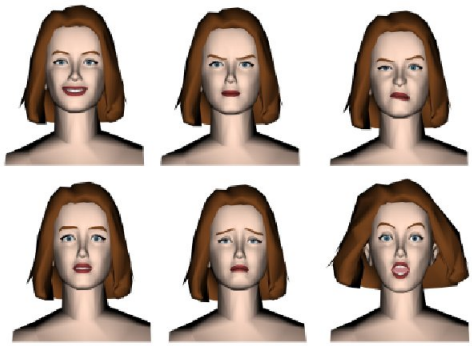


Figura 3.14: Expressões de alto nível do MPEG-6 FAP artístico.

Como o padrão MPEG-4 não enfatiza os detalhes do procedimento de animação, faz-se necessária uma forma de especificar como um FAP é interpretado de uma maneira mais precisa. Assim, o padrão MPEG-4 também especifica o uso de uma *Facial Animation Table* (FAT), que determina que vértices do modelo geométrico da face são afetados por um determinado FAP. Portanto, o modelo FAT é muito útil, garantindo não apenas o formato preciso da face mas também a reprodução exata da animação. Assim, se por um lado o motor de deformação do MPEG-4 clássico é baseado em algoritmos complexos, o motor MPEG-4 FAT é baseado em funções simples de interpolação. Por essa razão, este trabalho utiliza métodos de construção de FAT para aplicações para a *web* com alta precisão artística.

O motor MPEG-4 de animação facial trabalha com faces conformes, existindo diferentes metodologias a serem aplicadas. Um primeiro método consiste em definir a correspondência entre os vértices da malha da face e os pontos característicos do MPEG-4. Uma vez definidos esses dados, a geração da FAT se torna um procedimento fácil e que é feito através do programa desenvolvido no MIRALab [MT89], que faz uso do motor de animação facial do MPEG-4. Para cada FAP verifica-se a posição depois de passar por deformações comparando-a com sua posição em seu estado neutro, possibilitando assim a construção das tabelas de deformação para cada FAP e para cada vértice. Este método é baseado no motor de animação facial do MPEG-4 e pode ser utilizado por qualquer modelo facial compatível com MPEG-4. A grande vantagem deste método é a velocidade da construção da FAT, porém não é possível desenvolver uma deformação "manual" como a feita por um animador.

Um segundo método é através da construção da FAT artística. Quando um animador desenvolve um modelo que será animado por um conjunto de FAPs, a definição de uma posição neutra (estado de ânimo neutro) é importante, e será usada posteriormente para a animação. O

motor de animação gerencia diretamente as deformações para os FAPs relacionados à rotação. Geralmente, durante a animação, a construção de apenas um intervalo para cada FAP é suficiente, mas, se necessário, o animador pode usar vários intervalos. A vantagem de maior importância neste método é a capacidade de controlar a deformação facial de uma forma exata em termos da intensidade do FAP, porém este método possui um tempo de processamento bem maior do que o método descrito anteriormente.

Objetivando extrair os benefícios de ambos os métodos, é possível usar o motor do MPEG-4 do MIRALab para construir uma FAT para FAPs de baixo nível e usar um animador para a construção dos visemas e das expressões faciais, encontrando assim uma metodologia híbrida.

A implementação atual do *player* de animação facial do MPEG-4 por meio de FAT é escrita na linguagem de programação Java [Jav95] e utiliza a máquina de visualização Shout3D [Sho01], podendo ainda fazer uso do motor do OpenGL [Ope92], buscando uma melhoria no desempenho do sistema. A aplicação proposta através do trabalho *Designing MPEG-4 Facial Animation Tables for Web Applications* priorizou desenvolver o motor em Java para simplificar a integração na *web*, não fazendo uso de *plug-in*. Como a abordagem aqui apresentada é bastante modular, o uso da biblioteca desenvolvida, libFAT, é bastante simples. O primeiro passo consiste da inicialização do procedimento, capaz de carregar e compilar os dados da FAT no motor e fornecer outras informações sobre o modelo, como, por exemplo, sobre a posição neutra. Durante a animação, para cada quadro é carregado o conjunto de FAPs na libFAT. Após essa etapa, para cada parte constituinte da face diferente, a deformação da malha é computada pela libFAT através de uma função de atualização. Por fim, para que o personagem virtual seja mostrado, é necessário fazer o *download* de um conjunto de arquivos diferentes, sendo também necessário para a execução da aplicação fazer o *download* do modelo facial e da informação da FAT correspondente.

Uma vez definida uma aplicação *talking head* tomando como base o modelo de animação facial apresentado no trabalho *Designing MPEG-4 Facial Animation Tables for Web Applications* (padrão MPEG-4), é possível classificar esta aplicação segundo a taxonomia definida. A aplicação caracteriza-se pela obtenção da fala através de um sistema TtS e pela definição da face através de um modelo geométrico tridimensional. Quanto à forma de execução, o trabalho se propõe a ser um sistema em tempo real, mas esta classificação fica inerente à aplicação que está sendo desenvolvida. Por exemplo, a aplicação proposta no trabalho é de execução em *batch*, segundo a idéia de tempo real para sistemas *talking head*.

3.5 Taxonomia e Trabalhos Relacionados

A Tabela 3.1 apresenta um resumo da taxonomia definida neste capítulo, descrevendo as abordagens que cada parâmetro (fala, face e forma de execução) pode assumir.

Tabela 3.1: Taxonomia proposta para a classificação de sistemas *talking head*.

Parâmetros de Classificação	Abordagem 1	Abordagem 2
Fala	sintetizada	capturada
Face	imagens capturadas	malha poligonal
Forma de Execução	tempo real	*batch*

Com a taxonomia proposta e alguns dos trabalhos existentes apresentados, é possível cruzar

essas informações. O resultado deste cruzamento é a classificação de cada sistema segundo os parâmetros fala, face e forma de execução, como ilustra a Tabela 3.2 [1].

Tabela 3.2: Classificação dos trabalhos relacionados apresentados segundo a taxonomia proposta para sistemas *talking head*.

Trabalhos Relacionados	**Fala**	**Face**	**Forma de Execução**
Video Rewrite	capturada	imagens capturada	*batch*
MikeTalk	sintetizada	imagens capturada	*batch*
Facade	capturada	malha poligonal	*batch*
FaceWorks	capturada	malha poligonal com adição de texturas	*batch*
Padrão MPEG-4	capturada	híbrida	*batch*
Expressive Talking Heads	sintetizada	malha poligonal	*tempo real*

[1] A classificação do "Expressive Talking Heads" foi colocada na tabela apenas como fator comparativo em relação aos outros sistemas. Cada abordagem assumida para os parâmetros fala, face e forma de execução será justificada e apresentada detalhadamente nos próximos capítulos.

Capítulo 4

O Expressive Talking Heads

O "Expressive Talking Heads" é o resultado final de uma Dissertação de Mestrado que propõe a construção de uma aplicação de animação facial que possui como requisitos operacionais a interatividade, portabilidade e extensibilidade, e adicionalmente, os requisitos para aplicação *web* apresentados na Seção 2.3.3. O sistema desenvolvido objetiva estabelecer uma animação facial natural, buscando produzir em um personagem virtual comportamento semelhante ao da face humana durante a fala.

Este capítulo apresenta a estruturação do sistema "Expressive Talking Heads", descrevendo cada um de seus módulos. Em alguns casos, os módulos foram acoplados a subsistemas já existentes que foram incorporados ao "Expressive Talking Heads" para tornar o produto desta pesquisa mais rico em recursos oferecidos e mais adaptável a novas extensões. É importante ressaltar que este capítulo apresenta uma visão global do sistema, ficando para o Capítulo 5 a descrição dos detalhes de implementação. Ainda neste capítulo, são apresentadas as principais características de cada um dos subsistemas incorporados, como também a metodologia utilizada para a integração desses subsistemas no "Expressive Talking Heads".

4.1 Visão Geral do Expressive Talking Heads

O "Expressive Talking Heads" foi projetado para ser um sistema de animação facial interativo, capaz de interpretar as ações do usuário e refleti-las na animação final de um personagem virtual. Essas ações estão intimamente relacionadas com o discurso a ser falado e com a emoção que o personagem deve assumir para cada trecho da fala. Algumas outras características como gênero da voz e idioma da fala também foram previstos para as interações.

A partir desse objetivo, foi necessário pesquisar qual seria a melhor abordagem para cada um dos parâmetros apresentados no Capítulo 3. O primeiro parâmetro estudado foi a fala. Foram identificadas duas abordagens para o tratamento da fala: fala capturada e fala sintetizada (Seção 3.1). Para contemplar o requisito de interatividade, foi escolhida a segunda abordagem (fala sintetizada), onde o usuário interage continuamente com o sistema através de uma entrada textual via digitação ou leitura de um arquivo. O texto pode conter a fala e, adicionalmente, marcações sobre o estado de ânimo do personagem, o gênero (masculino ou feminino) da voz, o idioma/sotaque etc. O texto de entrada deve então passar por um subsistema sintetizador TtS (*Text-To-Speech*), sem impedir que o usuário continue a interagir com a aplicação. Três motivos levaram ao descarte da abordagem de fala capturada. O primeiro motivo foi a própria interface prevista para o sistema, onde ocorrendo a entrada de dados em paralelo com a animação, utilizando a abordagem de voz capturada acabaria resultando em

uma interferência de sons. O segundo motivo foi o aumento de complexidade e conseqüente ineficiência que poderia ocorrer com a incorporação de um subsistema de reconhecimento de voz na entrada do sistema, inviabilizando a interatividade de maneira continuada (tempo real). A terceira razão para a não utilização da fala capturada foi a questão do tratamento de emoção. A identificação da emoção já incorporada ao áudio seria uma tarefa complexa. Além disso, a utilização de falas previamente armazenadas dificultaria, ou mesmo impediria, a modificação do seu estado emocional. Entretanto, nada impede que, para novas aplicações que venham a ser derivadas do "Expressive Talking Heads" seja investigada a possibilidade de incorporar ao ambiente um subsistema de reconhecimento de voz, possibilitando dessa forma a operação nas duas abordagens.

O segundo parâmetro analisado foi a face. A definição da face consiste de duas etapas: a modelagem e a sua posterior animação. Como visto na Seção 3.2, a modelagem da face pode assumir duas abordagens durante sua construção: face definida através de imagens capturadas e face definida através de uma malha poligonal. Na face definida através de imagens capturadas, o usuário deve ficar fornecendo as imagens a serem usadas para compor a animação facial. A transição de uma imagem para outra é feita através da técnica de metamorfose, não constituindo assim uma boa abordagem para o tratamento de sistemas *talking head* interativos. Por outro lado, na abordagem de definição da face através de uma malha poligonal, o usuário fornece o texto e a descrição da emoção da fala do personagem e, a partir daí, torna-se responsabilidade do sistema interpretar a emoção e os visemas que compõem aquela fala para executar a animação facial sobre os vértices da malha poligonal, através de uma das técnicas apresentadas na Seção 2.3.2. A partir da análise comparativa entre essas duas abordagens, foi possível verificar que a construção de um sistema *talking head* onde a face é definida através de um modelo poligonal atende melhor ao requisito de interatividade. Mais ainda, as faces geradas por imagens capturadas, embora mais realistas, sofrem com a falta de expressividade, como é evidenciado pelo trabalho desenvolvido no sistema MikeTalk [EP99] [EP98].

O terceiro parâmetro analisado foi a forma de execução do sistema (Seção 3.3). A forma de execução assumida para o "Expressive Talking Heads" foi a execução em tempo real. Essa abordagem foi escolhida pois, se o sistema busca um maior grau de interatividade, ele se tornaria uma ferramenta mais poderosa se as interações que o usuário fizesse com o sistema fossem refletidas imediatamente após sua definição. Na realidade, essa abordagem foi escolhida como um pré-requisito para o sistema e, conforme pôde ser percebido com as considerações anteriores, a definição dessa abordagem direcionou a escolha dos demais parâmetros.

Portanto, é possível definir o "Expressive Talking Heads" como um sistema *talking head* que recebe um texto como entrada para gerar como saída uma animação facial de um personagem virtual enunciando o texto fornecido. O texto de entrada deve conter a fala a ser enunciada e pode incluir informações sobre o estado de ânimo do personagem. A entrada textual é interpretada e o texto propriamente dito é enviado para o sintetizador, gerando assim o sinal de áudio equivalente à fala. A partir das informações sobre o estado de ânimo do personagem e com a fala sintetizada, é gerada uma animação sobre o modelo poligonal estabelecido para a face. Essa animação atua sobre os vértices da malha que funcionam como pseudo-músculos. Por fim, todo este processo é executado de forma interativa e em tempo real, ou seja, uma vez enviado um texto e iniciado o seu processamento, o usuário já pode novamente enviar uma nova fala para o personagem virtual. A próxima seção apresenta de forma resumida a organização dos módulos no "Expressive Talking Heads", enquanto as seções seguintes discorrem sobre os subsistemas já existentes que foram adaptados ao ambiente desenvolvido.

4.1.1 Os Módulos do Expressive Talking Heads

Objetivando construir o sistema através de uma abordagem que permitisse reuso, o "Expressive Talking Heads" foi estruturado como sendo composto por três módulos: módulo de síntese da entrada, módulo de gerenciamento da face e módulo de sincronização. A Figura 4.1 oferece uma visão geral da estruturação do "Expressive Talking Heads" com seus principais módulos.

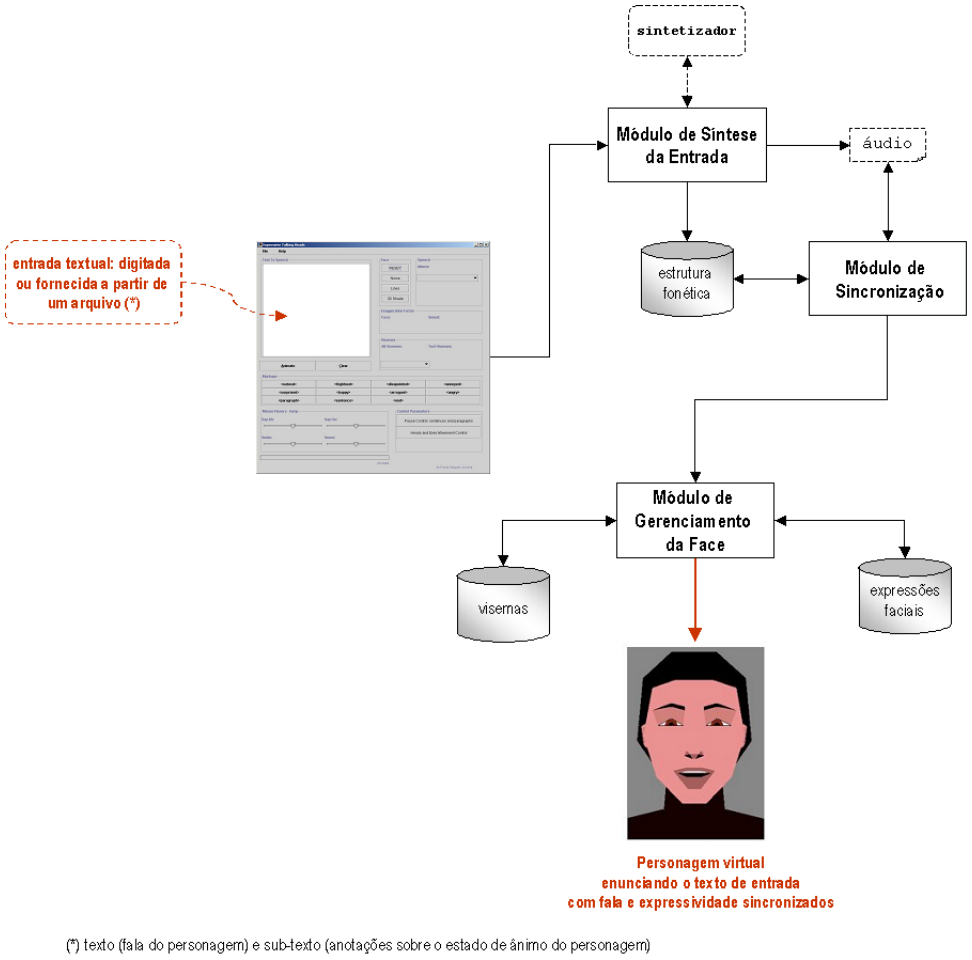


Figura 4.1: Visão Geral do Expressive Talking Heads.

O módulo de *síntese da entrada* é responsável por capturar a entrada textual e interpretá-la, separando a fala propriamente dita das marcações sobre o estado de ânimo do personagem. O texto contendo a fala é enviado para o sintetizador TtS, que retorna como saída as informações fonéticas do texto e o áudio contendo a fala.

O segundo módulo do "Expressive Talking Heads", o módulo de *gerenciamento da face*, mantém a base de visemas e a base de expressões faciais. Este módulo aguarda o início da animação facial para começar a manipular os componentes faciais em sincronia com a fala. A manipulação dos componentes faciais é realizada pela aplicação dos valores de contração e relaxamento sobre os vértices da malha poligonal. Esses valores que alteram o posicionamento dos músculos da face são extraídos tanto dos visemas como da expressão facial, sempre tomando por base o fonema que identifica cada instante da fala. Na realidade, a expressão facial e o visema são gerados a partir da informação dos difones, buscando com isso preservar na animação os aspectos de coarticulação.

O "Expressive Talking Heads" é composto ainda pelo módulo de *sincronização*. Este

módulo recebe como entrada a saída proveniente do módulo de síntese da entrada, sendo sua responsabilidade gerar uma animação facial com o conteúdo dos dois outros módulos sincronizados. Uma vez que o áudio com a fala do personagem virtual começa a ser apresentado, o módulo de sincronização verifica qual o fonema e a emoção que estão associados àquele trecho de fala, buscando o visema e a emoção correspondentes de forma que a transição do estado atual seja bastante suave, gerando assim a animação facial desejada.

Como já mencionado, o Capítulo 5 destina-se a apresentar cada um destes módulos de forma detalhada, descrevendo seus principais aspectos de implementação.

4.2 Subsistemas que Compõem o Expressive Talking Heads

Cada um dos três módulos do "Expressive Talking Heads" possui uma arquitetura própria e funcionalidades bem definidas. Além disso, cada módulo pode ser dividido em submódulos e, em alguns casos, fazer uso dos serviços oferecidos por subsistemas. Esses subsistemas são aplicações que já existiam e que foram incorporadas a esse projeto de pesquisa.

Esta seção destina-se a apresentar os três subsistemas externos incorporados ao "Expressive Talking Heads": o Festival, o MBROLA e o Responsive Face. São abordados apenas aspectos conceituais de cada subsistema, ficando para o Capítulo 5 a apresentação dos detalhes de implementação para suas incorporações assim como as dificuldades encontradas nesse processo.

4.2.1 O *Festival Speech Synthesis System*

O primeiro subsistema a ser incorporados ao "Expressive Talking Heads" foi o *Festival Speech Synthesis System* [WT97], um sintetizador de textos desenvolvido na Universidade de Edimburgo, com distribuição gratuita e sem fins comerciais.

O Festival é implementado em duas linguagens, C++ e *Scheme* [JS] (uma variação de Lisp). O uso das duas linguagens na construção deste sintetizador decorre da arquitetura complexa do sistema [WTC99a] e da intenção de explorar as particularidades e os benefícios de uma linguagem de baixo nível (C++) e de uma linguagem baseada em *script* (*Scheme*).

O uso de uma linguagem de baixo nível como C++ permitiu que o Festival operasse também como um sistema de execução, além de uma plataforma de pesquisa, sendo o desempenho um requisito fundamental. Por outro lado, o uso de uma linguagem interpretada como *Scheme* permitiu que se fizessem modificações em tempo de execução nos *scripts*, facilitando a interação do usuário com o sistema, sem necessidade de alterar o código de baixo nível do sintetizador e recompilá-lo para que a nova execução refletisse as modificações/extensões.

No Festival existe uma distinção entre o núcleo do sistema e os módulos que executam as tarefas de síntese da fala. O núcleo do sistema define a arquitetura e é completamente escrito em C++, não podendo ser modificado. Por outro lado, os módulos podem ser escritos tanto em C++ como em *Scheme* e podem ser adicionados ou alterados sem afetar a base do sintetizador.

Como ocorre com a maioria dos sistemas TtS, o Festival divide o problema de converter o texto em fala em duas sub-tarefas: na primeira, a unidade NLP (processamento da linguagem natural) converte o texto de entrada em um conjunto de informações relevantes à fonética, duração e parâmetros de entonação. Já na segunda etapa, a unidade DSP (processamento do sinal digital) converte a saída da unidade NLP em um áudio que enuncia o texto de entrada.

O Festival foi projetado como um sistema de síntese da fala que pode ser utilizado por usuários de três diferentes níveis. O primeiro, são aqueles usuários que utilizam o sistema apenas como um sintetizador de fala de alta qualidade, síntese que é realizada a partir de uma entrada textual arbitrária, com um esforço mínimo por parte do usuário. O segundo nível engloba aqueles usuários que desenvolvem sistemas de linguagens que desejam incluir saídas sintetizadas. Neste caso, algumas configurações personalizadas são desejadas, como vozes diferentes, expressões específicas, tipos de diálogo, entre outras. O terceiro e último nível possibilita utilizar o Festival para o desenvolvimento e teste de novos métodos de síntese.

O "Expressive Talking Heads" utiliza o Festival como um sintetizador de fala a partir de entradas textuais, atuando portanto como um usuário de primeiro nível. Na realidade, conforme ficará mais claro com a apresentação do projeto MBROLA, em algumas situações o "Expressive Talking Heads" explora configurações personalizadas, como diferentes vozes e idiomas, junto ao sintetizador. No entanto, quem efetivamente atua como um usuário de segundo nível do Festival é o MBROLA, permanecendo o "Expressive Talking Heads" como um usuário de primeiro nível do sintetizar combinado Festival-MBROLA.

Além de possuir três níveis de interação com o usuário, o Festival pode ser utilizado dentro de alguma outra aplicação, disponibilizando para isto diversas interfaces de programação (APIs) [WT97]. É desta maneira que o "Expressive Talking Heads" incorpora o Festival como um subsistema e utiliza os seus serviços de síntese.

O "Expressive Talking Heads" baseia-se na API cliente-servidor do Festival, na qual o Festival executa como servidor e o "Expressive Talking Heads" como cliente. A comunicação é feita através de uma conexão TCP, permitindo que os dois elementos se encontrem em máquinas distintas. Esse modelo distribuído foi especialmente interessante para o desenvolvimento da aplicação como um *applet*, conforme será discutido no Capítulo 5. Os comandos passados para o Festival a partir do "Expressive Talking Heads" são codificados na linguagem *Scheme*.

Por fim, o Festival é um sistema multi-língua, trabalhando com os idiomas inglês (americano e britânico), espanhol e gales, sendo para o inglês sua implementação mais avançada (em termos de vocabulário). O "Expressive Talking Heads" faz uso desta plataforma multi-língua do Festival em conjunto com um outro sintetizador (Seção 4.2.2 e Seção 5.1.3). O Capítulo 5 discute os detalhes de implementação para integração do Festival ao "Expressive Talking Heads".

4.2.2 O Projeto MBROLA

O segundo subsistema que compõe o "Expressive Talking Heads" é o *MBROLA* (*Multi-Band Resynthesis Overlap-Add speech synthesis*) [Dea97] [Dea96], sistema inicialmente desenvolvido pelo TCTS Lab da Faculté Polytechnique de Mons (Bélgica). O objetivo do projeto MBROLA é obter um conjunto de sintetizadores de fala para o maior número possível de vozes, idiomas e dialetos, também sendo de distribuição gratuita. O principal objetivo é incentivar a pesquisa acadêmica na área de síntese da fala, em particular na geração de prosódia, tida como um dos maiores desafios dos sintetizadores *Text-to-Speech* até os dias de hoje.

O MBROLA é um sintetizador de fala baseado na concatenação de difones. A entrada é uma lista de alofones [1] associados com informações da prosódia (duração e entonação). A saída produzida pelo sintetizador consiste de amostras lineares de fala de 16 *bits*. Ao contrário do Festival, o MBROLA não é um sistema *Text-to-Speech*, pois não aceita como entrada um

[1] Entendem-se por *alofones* as variações para um determinado fonema. Por exemplo, para o /t/ há as seguintes variações de realizações fonéticas: *ta-te-ti-to-tu* [Bec01].

"texto natural".

O MBROLA possui uma base de dados de fonemas padrão e um conjunto de mapeamentos desses fonemas para os fonemas em outros idiomas. Atualmente, ele trabalha com 17 idiomas possíveis, incluindo inglês (americano e britânico), português, francês e alemão, entre outros. Além disso, o MBROLA permite que a voz seja personalizada - por exemplo, atualmente é possível escolher entre uma voz feminina ou masculina.

O "Expressive Talking Heads" incorporou o MBROLA ao seu módulo de síntese da entrada. Um dos principais motivos foi a plataforma multi-língua que o MBROLA oferece, favorecendo o enriquecimento do sistema. Porém, como o MBROLA não recebe "textos naturais" como entrada, ele foi utilizado em conjunto com o Festival para produzir a síntese de um texto de entrada em diferentes idiomas e gêneros. Esta integração Festival-MBROLA é uma das contribuições deste trabalho e será apresentada em detalhes na Seção 5.1.3.

A integração dos dois subsistemas faz com que o MBROLA atue no "Expressive Talking Heads" como a unidade DSP (processamento do sinal digital), enquanto o Festival atue apenas como a unidade NLP (processamento da linguagem natural). A saída do Festival (NLP), que consiste de informações fonéticas (alofones com duração e entonação), é a entrada para o sintetizador MBROLA (DSP). Esse, por sua vez, gera como saída o áudio digitalizado, completando assim o processo de síntese.

A integração dessas duas ferramentas ao sistema permitiu também que o "Expressive Talking Heads" interferisse no processo de síntese, buscando colocar um maior realismo na saída de áudio gerada. A idéia foi interceptar e manipular a saída fonética do Festival, basicamente modificando o tratamento de pausa da unidade NLP, e repassando essa descrição modificada para o MBROLA a fim de refletir as alterações na saída de áudio gerada. Esse processo é melhor explicado no próximo capítulo. Um possível trabalho futuro, seria estender essa manipulação aos parâmetros de entonação e duração dos demais fonemas, principalmente para ter um maior controle sobre a emoção na voz sintetizada.

4.2.3 O *Responsive Face*

O terceiro subsistema que compõe o "Expressive Talking Heads" é o Responsive Face [Per97], desenvolvido por Ken Perlin no Media Research Lab, na Universidade de Nova York. Como o próprio nome sugere, o Responsive Face teve como principal objetivo a construção de um sistema interativo com o personagem virtual capaz de formular uma comunicação face a face com o usuário. O Responsive Face trabalha com um número mínimo de elementos de expressão facial que permite produzir uma impressão convincente do personagem e de sua personalidade e emoção.

O sistema tem como requisitos apresentar-se interativo e dinâmico, possibilitando a combinação das expressões faciais definidas a fim de simular variações do estado de humor e atitudes do personagem virtual. De modo simplificado, é possível definir o Responsive Face como um sistema interativo de animação facial com expressividades emotivas e convincentes.

O objetivo visado com essa pesquisa e com o desenvolvimento desse sistema é permitir que interfaces homem-computador sejam capazes de representar as sutilezas que são próprias da comunicação face a face, de forma que as interfaces possam funcionar como agentes de um ponto de vista emocional. No Responsive Face busca-se, a partir de um número mínimo de expressões faciais, produzir uma impressão convincente do personagem e sua personalidade. É evidente que o conjunto de expressões utilizado é apenas um subconjunto do intervalo completo de expressões que uma face humana é capaz de apresentar.

O Responsive Face também é resultado de um desafio pessoal do autor Ken Perlin, que demonstrou que poderia implementar um personagem tridimensional na linguagem de programação Java [Jav95], sem utilizar qualquer *plug-in 3D*, isto é, obtendo a visualização inteiramente no próprio *applet* Java. No entanto, é importante salientar que nenhum tratamento à fala foi incorporado ao personagem no projeto original do sistema.

Sendo assim, o "Expressive Talking Heads" aproveitou do Responsive Face a face e as expressões faciais já modeladas e adicionou fala a esse personagem virtual. Evidentemente, o sistema desenvolvido cuidou de estabelecer o sincronismo entre a fala e os movimentos faciais, incluindo na animação as expressões para emoções que já haviam sido definidos no Responsive Face. A Seção 5.2 discute os detalhes dessa integração.

4.3 Integração dos Subsistemas em um Sistema *Talking Head*

Uma vez conhecendo os principais módulos e subsistemas que compõem o "Expressive Talking Heads", o próximo passo consiste em integrá-los para permitir a comunicação e o funcionamento do sistema como um todo. A partir da Figura 4.1 é possível identificar os submódulos do sistema: o submódulo de captura dos dados de entrada, o submódulo de *parse*, o submódulo de gerência da síntese, o submódulo de visemas e o submódulo de tratamento das emoções. É importante ressaltar ainda que os próprios módulos possuem funcionalidades definidas para estabelecer esta integração.

O submódulo de captura dos dados de entrada tem como objetivo capturar os valores dos parâmetros definidos na interface gráfica do sistema pelo usuário. Esses parâmetros transmitem informações sobre o texto que deverá compor a fala do personagem e o respectivo estado de ânimo. Recursos de configuração de idioma e do gênero da voz também foram incluídos. A saída deste submódulo será a entrada para o submódulo de *parse*.

O *parser* é responsável por identificar e separar as informações provenientes do primeiro submódulo. Uma vez analisadas, as informações são entregues ao submódulo de gerência da síntese. A idéia é que o *parser* informa, trecho a trecho, o texto da fala, já indicando quais são os parâmetros de emoção, idioma/sotaque, gênero para o trecho.

O submódulo de gerência da síntese utiliza os serviços dos subsistemas de síntese TtS (Festival e MBROLA) e constrói uma base de dados contendo a descrição da animação. Basicamente, essa base irá possuir a relação de fonemas, duração, entonação e a emoção para cada um dos fonemas. Esses dados serão utilizados pelos módulos de sincronização e de gerenciamento da face para construir as expressões faciais sincronizadas com a fala.

Os submódulos de visemas e de tratamento de expressões são os principais componentes que formam o módulo de gerenciamento da face. O submódulo de visemas é responsável por construir a base de visemas do sistema. Este submódulo também realiza a transição de um visema para o seu subseqüente, a partir de informações provenientes do arquivo de fonemas, como será melhor detalhado posteriormente. Já o submódulo de tratamento de expressões recebe informações sobre o estado de ânimo do personagem e sabe como definir o mapeamento para as expressões da face. A partir de uma dada expressão, o submódulo identifica quais os músculos faciais afetados (os vértices da malha poligonal do *Responsive Face*) e quais os valores de contração/relaxamento que devem ser aplicados.

Capítulo 5

Aspectos de Implementação do Expressive Talking Heads

Este capítulo destina-se a apresentar os aspectos de implementação do sistema "Expressive Talking Heads". Como já mencionado, o objetivo desta aplicação é receber como entrada um texto contendo a fala e anotações e gerar como saída, em tempo real, a animação de um personagem virtual enunciando o texto de entrada com o áudio e os movimentos faciais sincronizados.

O sistema foi estruturado em três módulos: *síntese da entrada*, *gerenciamento da face* e *sincronização*. Cada uma das três primeiras seções deste capítulo apresenta um desses módulos, descrevendo a visão geral do módulo, os elementos que o constituem, suas importâncias, participações e principais detalhes de implementação.

O sistema "Expressive Talking Heads" foi desenvolvido utilizando a linguagem de programação Java [Jav95] em duas versões: uma aplicação de execução local (*stand alone*) e uma aplicação para execução em navegadores *web* (*applet*). A principal razão da escolha de Java foi justamente favorecer a execução em ambientes WWW (*World Wide Web*), através da portabilidade oferecida pelo modelo de *bytecode* da linguagem. A Seção 5.4 destina-se a descrever as adaptações do sistema para as duas formas de execução. Por fim, para o leitor que tiver interesse em maiores detalhes da implementação, o Apêndice A apresenta uma descrição completa dos diagramas de classe do sistema.

5.1 Módulo de Síntese da Entrada

A Figura 5.1 apresenta a visão geral do módulo de síntese da entrada. Esse módulo é responsável por capturar e tratar o texto fornecido como entrada pelo usuário e gerar como saída uma estrutura de dados contendo as unidades fundamentais para a geração da animação facial (fonema, duração, emoção etc.) e o áudio digitalizado da fala correspondente ao texto de entrada.

O texto de entrada é fornecido através de uma linguagem de marcação contendo informações sobre a emoção do personagem, o gênero da voz (atualmente masculino e feminino adultos) e o idioma da fala (atualmente inglês americano e inglês britânico). Esse texto pode ser fornecido através da digitação do usuário ou de um arquivo previamente definido e armazenado. O módulo de síntese interpreta o texto marcado, através de um elemento *parser*, separando o conteúdo da fala propriamente dita das informações de controle. A partir do texto e das marcações analisadas pelo *parser*, o *sintetizador ETHs* faz uso dos serviços oferecidos

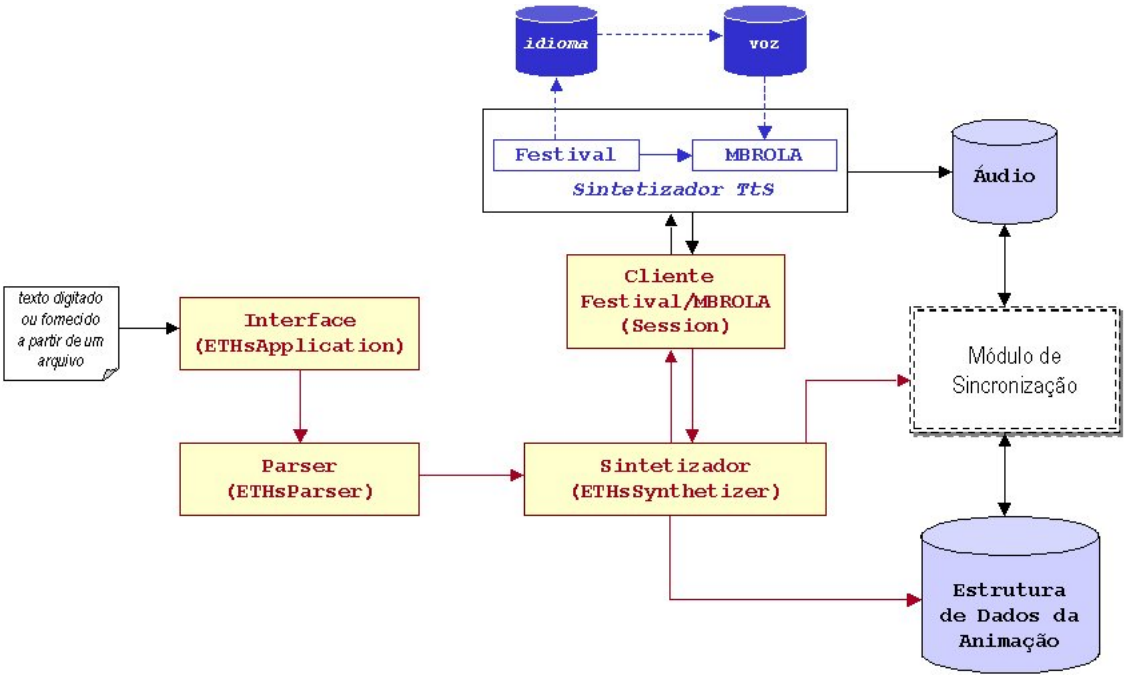


Figura 5.1: Visão geral do módulo de síntese da entrada. Os nomes entre parênteses correspondem aos nomes das classes Java implementadas no sistema.

pelos subsistemas Festival e MBROLA para obter a descrição fonética e o áudio digitalizado correspondentes à fala do personagem. Nessa etapa, dependendo da opção selecionada pelo usuário na interface do sistema, o *sintetizador* acrescenta um tratamento especial à pausa entre as sentenças da fala, a fim de melhorar o realismo do áudio gerado.

Uma vez que o módulo de síntese da entrada é o módulo destinado a lidar diretamente com as ações do usuário, em função da entrada dos dados e da opção de interatividade escolhida pelo usuário da aplicação, este módulo coordena as ações do módulo de sincronização. É exatamente essa relação entre os módulos de síntese e de sincronização que confere a característica de interatividade em tempo real do "Expressive Talking Heads".

O restante desta seção aborda em detalhes o funcionamento e os relacionamentos dos componentes do módulo de síntese da entrada ilustrados na Figura 5.1. O primeiro elemento a ser abordado é a unidade fundamental utilizada na construção da estrutura de dados que contém a descrição da animação facial. Essa estrutura servirá como entrada para os módulos de sincronização e de gerenciamento da face.

5.1.1 Unidades Fundamentais

O módulo de síntese da entrada é composto por uma única unidade fundamental, o *fonema* (classe *Phoneme*). A estrutura de um fonema no "Expressive Talking Heads" é caracterizada por cinco atributos internos: rótulo, duração, tempo esperado para ocorrer, entonação e expressão facial (Figura 5.2). Os fonemas (objetos da classe *Phoneme*) são fundamentais no "Expressive Talking Heads", sendo instanciados no módulo de síntese da entrada e utilizados nos outros módulos do sistema.

O *rótulo* do fonema é o seu identificador. Este atributo é inicializado pelo sintetizador TtS e, uma vez criado, não sofre mais alteração. Cada idioma possui um conjunto de fonemas próprio. Além disso, os identificadores definidos pelo sintetizador TtS para um mesmo fonema podem variar de um idioma para outro. Por exemplo, para o inglês americano, o Festival

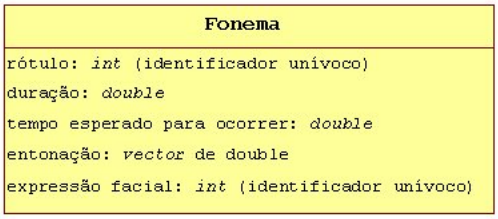


Figura 5.2: O elemento *fonema*.

representa o fonema do silêncio através da *string* "*pau*", enquanto para o inglês britânico este fonema é representado pela *string* "#".

O atributo *duração*, assim como o rótulo, é inicializado pelo sintetizador TtS e corresponde ao tempo em milisegundos que cada fonema leva para ser pronunciado. Este elemento também varia de acordo com o idioma, mas, diferente do rótulo, ele pode ser alterado pelo sintetizador ETHs [1]. Por exemplo, os fonemas de pausa ("*pau*") são sempre gerados pelo Festival com uma duração de 220 milisegundos, mas em alguns modos de operação do sistema essa duração é alterada pelo sintetizador ETHs para dar um maior realismo à fala. O tratamento da pausa será abordado em maiores detalhes na Seção 5.1.4.

O atributo *tempo esperado para ocorrer* é um valor numérico inicializado pelo sintetizador ETHs, sendo derivado das durações dos demais fonemas contidos no texto da fala. O tempo esperado para ocorrer de um determinado fonema é determinado através do somatório das durações dos fonemas anteriores, representando o instante esperado para que seja iniciada a pronúncia do fonema. A principal razão de armazenar o tempo esperado na estrutura é tornar mais eficiente a identificação do fonema corrente durante a sincronização labial.

O atributo *entonação* está relacionado com a prosódia do fonema durante a fala. A entonação também é gerada automaticamente pelo sintetizador TtS e pode ser modificada, principalmente quando a expressividade é levada em consideração. Por exemplo, o ritmo e o timbre (tom) da fala de uma pessoa são diferentes quando se está triste ou feliz. É justamente sobre este conjunto de valores que estas alterações se refletem. Neste trabalho, a entonação dos fonemas não foi tratada, sendo sempre utilizados os valores *default* inicializados pelo sintetizador TtS. No entanto, a estrutura permite que em um trabalho futuro esse tratamento seja feito para dar um maior realismo à saída, sem que a estrutura de dados precise ser alterada.

O quinto e último atributo de um fonema é a expressão facial, representada por um identificador numérico para cada possível estado de ânimo do personagem. Diferente dos atributos rótulo, duração e entonação, que são inicializados pelo sintetizador TtS, este atributo é inicializado diretamente pelo sintetizador ETHs, com o valor informado pelo *parser*. A próxima seção discute o tratamento do texto de entrada, apresentando a linguagem de marcação definida e comentando os tipos de expressão facial considerados pelo "Expressive Talking Heads".

5.1.2 Interpretando o Texto de Entrada

O tratamento do texto de entrada no "Expressive Talking Heads" é responsabilidade de um elemento denominado *parser ETHs* (*ETHsParser*), que faz a ponte entre o texto de entrada fornecido pelo usuário e o sintetizador ETHs (*ETHsSynthetizer*). A Figura 5.3 ilustra os componentes e relacionamentos do submódulo de *parser*.

[1] Abreviação assumida para "Expressive Talking Heads".

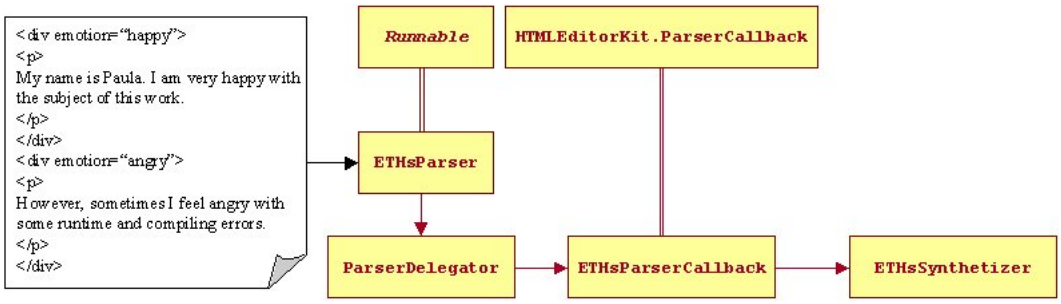


Figura 5.3: Visão geral do *ETHsParser*.

A linguagem de marcação definida para a entrada dos dados foi uma pequena extensão da linguagem HTML [W3C00]. A princípio, foi cogitada a criação de uma nova linguagem de marcação baseada no padrão XML [W3C99]. No entanto, essa abordagem obrigaria a utilização de um pacote externo ao pacote *default* da linguagem Java 2. Para o caso da implementação da aplicação como *applet*, isso tornaria necessária a instalação prévia de um pacote na máquina do cliente ou o aumento em quase um *megabyte* na quantidade de dados a ser transmitida na rede quando fosse feito o acesso ao sistema.

Por essas razões, este trabalho de pesquisa optou por estender a linguagem HTML, aproveitando o suporte para a análise desses tipos de documentos existente no pacote *javax.swing.text.html.parser* (classes *ParserDelegator* e *HTMLEditorKit.ParserCallback*). Esse pacote, por ser interno à linguagem Java em sua versão 2, dispensa a instalação ou o *download* em clientes cuja máquina virtual implemente essa versão da linguagem. Com a tendência crescente dos navegadores oferecerem suporte a linguagens baseadas na metalinguagem XML, pode se considerar como trabalho futuro a utilização de uma linguagem de marcação de propósito específico baseada nesse padrão. Nesse sentido, uma alternativa é investigar o uso da linguagem SSML (*Speech Synthesis Markup Language*), um trabalho atualmente em andamento no W3C [2].

A extensão da linguagem HTML definida neste trabalho prevê marcações para definir o estado de ânimo do personagem, o gênero da voz e o idioma da fala. Essas informações são especificadas como atributos do elemento *div*, já existente na linguagem. Portanto, as extensões são os nomes dos atributos para o elemento; no caso, *emotion* para emoção, *genus* para o gênero e *language* para o idioma. A Tabela 5.1 descreve a marcação para emoções, salientando os valores atualmente válidos para o sistema. Para qualquer valor inválido, o sistema assume a expressão natural (estado de neutralidade) como valor *default*. Da mesma forma, se o usuário não informar um estado de ânimo, a expressão natural também é assumida como valor *default*. Mais ainda, qualquer texto digitado fora de um elemento *div* é sintetizado utilizando os valores que foram estabelecidos como *default*.

A Tabela 5.2 descreve os atributos válidos para a marcação de gênero, e a Tabela 5.3 apresenta os valores para o idioma. No primeiro caso, o valor *default* do sistema é a voz feminina, enquanto para o idioma o sistema define o inglês americano como *default*.

A terceira coluna das tabelas apresenta representações alternativas das marcações na forma de um novo elemento (nova *tag*) para a linguagem HTML. Essas alternativas foram definidas na interface de entrada do sistema para simplificar a tarefa de digitação do usuário. Internamente, antes de fornecer os dados ao *parser*, o sistema substitui esses elementos alternativas pelas marcações utilizando o elemento *div*. Essa substituição é importante de ser feita porque o *parser* HTML utilizado (*HTMLEditorKit.ParserCallback*) simplesmente ignora elementos que não

[2] *World Wide Web Consortium* em http://www.w3.org.

Tabela 5.1: Valores para a marcação de estado de ânimo definidos no "Expressive Talking Heads".

Marcação	**Significado** (estado de ânimo)	**Representação Macro**
* *	natural	*<natural>*
* *	com medo	*<frightned>*
* *	decepcionado	*<disappointed>*
* *	aborrecido	*<annoyed>*
* *	surpreso	*<surprised>*
* *	feliz	*<happy>*
* *	arrogante	*<arrogant>*
* *	com raiva	*<angry>*

Tabela 5.2: Valores para a marcação de gênero definidos no "Expressive Talking Heads".

Marcação	**Significado**	**Representação Macro**
* *	masculino	*<male>*
* *	feminino	*<female>*

façam parte da linguagem, impedindo que eles sejam tratados. Já os atributos dos elementos, mesmo quando possuindo nomes estranhos à linguagem, são considerados e podem com isso ser analisados.

O *parser* no "Expressive Talking Heads" é implementado como um *thread* [Lea96] [OW97] separada no sistema. Isso permite realizar uma síntese e gerar uma animação facial paralelamente à entrada dos dados por parte do usuário, procurando garantir o requisito de interatividade do sistema.

Conforme já mencionado, para realizar o tratamento das marcações o *parser* do sistema faz uso do suporte oferecido pela própria linguagem Java. Um objeto da classe *ETHsParser* instancia um objeto da classe *ParserDelegator* e chama o método *parse* dessa classe, passando como parâmetros uma referência para o texto a ser analisado e outra para um objeto que funcionará como *callback* da análise. A classe *ETHsParserCallback* oferece justamente esse suporte ao tratamento da análise, tendo sido definida como uma subclasse da classe *HTMLEditorKit.ParserCallback* do próprio pacote da linguagem Java. Os métodos *handleStartTag*, *handleText* e *handleEndTag* foram implementados para tratar as marcações e direcionar os pedidos de síntese ao sintetizador ETHs (*ETHsSynthetizer*).

Tabela 5.3: Valores para a marcação de idioma definidos no "Expressive Talking Heads".

Marcação	**Significado**	**Representação Macro**
* *	inglês americano	*<en>*
* *	inglês britânico	*<engb>*

5.1.3 Festival e MBROLA: Trabalhando Juntos

A entrada do "Expressive Talking Heads" é um texto fornecido pelo usuário. É responsabilidade da aplicação capturar esse texto e enviá-lo para o sintetizador TtS. No "Expressive Talking Heads" o papel de sintetizador TtS é desempenhado pelo trabalho conjunto de dois sintetizadores: o Festival (Seção 4.2.1 e [WT97]) e o MBROLA(Seção 4.2.2, [Dea97] e [Dea96]). Estes dois sistemas geram como saída informações fonéticas do texto de entrada e um áudio que enuncia o texto com as características estipuladas.

No funcionamento em conjunto, o Festival desempenha o papel de unidade de processamento da linguagem natural e o MBROLA exerce a função de unidade de processamento do sinal digital. A vantagem de utilizá-los juntos é a obtenção de um sintetizador TtS que oferece uma plataforma multi-língua e multi-personalizada (contribuição do MBROLA). Essa flexibilidade é importante para o "Expressive Talking Heads" porque permite que o idioma e o tipo da voz (gênero) sejam parâmetros escolhidos pelo usuário na síntese do texto, enriquecendo o sistema com a exploração dessas e outras características da fala expressiva.

O Festival-MBROLA é utilizado no sistema executando no modo servidor. O "Expressive Talking Heads" estabelece uma conexão TCP com o sintetizador TtS utilizando um objeto da classe *Session*, oferecida para a implementação de aplicações cliente do sintetizador TtS em Java [3]. Através dessa conexão, o "Expressive Talking Heads" envia comandos para o sintetizador efetuando chamadas ao método *synchronousRequest* e recebe as respostas encapsuladas em objetos genéricos Java (classe *Object*). As respostas do Festival-MBROLA são extraídas fazendo chamadas ao método *toString* do objeto retornado. Optou-se por utilizar a chamada síncrona da classe *Session* (método *synchronousRequest*) para aguardar o final da síntese de um determinado texto antes de iniciar o processo em outro. Essa abordagem facilitou o controle sobre o comportamento da execução, evitando problemas de concorrência nesse processo. A comunicação com o objeto da classe *Session* é coordenada pelo sintetizador ETHs (*ETHsSynthetizer*) - componente do módulo de síntese da entrada.

Os comandos enviados ao Festival devem ser codificados na linguagem de programação *Scheme* [JS]. Como de um modo geral as solicitações de síntese envolvem vários comandos em *Scheme*, foram desenvolvidas funções nessa linguagem e colocadas junto ao servidor Festival-MBROLA. Desse modo, o comando enviado ao servidor é apenas uma chamada à função, reduzindo a quantidade de dados transmitida na rede. As funções são então interpretadas e executadas na máquina que hospeda o servidor Festival-MBROLA.

A Função 5.1 descreve o procedimento *Scheme Synth_Pho_Ram*. Esse procedimento realiza o processamento NLP de um texto previamente definido em uma variável interna ao servidor Festival-MBROLA e armazena o resultado em uma outra variável interna, denominada *phoneme_structure*. Essa variável contém uma descrição fonética do texto em um formato como o apresentado na Figura 5.4. Antes de realizar uma chamada ao procedimento *Synth_Pho_Ram*, o sintetizador ETHs envia ao Festival-MBROLA o comando "*(set! text_to_speech "texto a ser sintetizado")*" para configurar o texto a ser sintetizado. Internamente, o procedimento *Scheme Synth_Pho_Ram* faz uso do procedimento *Save_Seg_Mbrola_Entry_Ram*, ilustrado na Função 5.2. Esse procedimento é chamado para cada fonema que precisa ser gravado na estrutura.

Após o término da interpretação do procedimento *Scheme Synth_Pho_Ram*, o sintetizador ETHs envia ao Festival-MBROLA, através de uma nova chamada ao método *synchronousRequest* do objeto da classe *Session*, o comando "(*phoneme_structure*)". Essa chamada é real-

[3] A porta *default* que o Festival (servidor) utiliza para aceitar conexões TCP é a 1314, podendo ser modificada se desejado.

Função 5.1 Sintetiza o texto de entrada armazenando a estrutura fonética em memória.

```
(define (Synth_Pho_Ram utt)``[Synth_Pho_Ram: sintetiza o texto em memória]''
   (set! phoneme_structure "")
     (mapcar
       (lambda (segment)
         (set! phoneme_structure
           (string-append phoneme_structure
             (Save_Seg_Mbrola_Entry_Ram
               (item.feat segment 'name)
               (item.feat segment 'segment_start)

               (item.feat segment 'segment_duration)
               (mapcar
                 (lambda (targ_item)
                   (list
                     (item.feat targ_item "pos")
                     (item.feat targ_item "f0")))
                 (item.relation.daughters segment 'Target)))) ;; list of targets
           )
         )
        (utt.relation.items utt 'Segment))
)
```

Função 5.2 Cria a estrutura de um fonema.

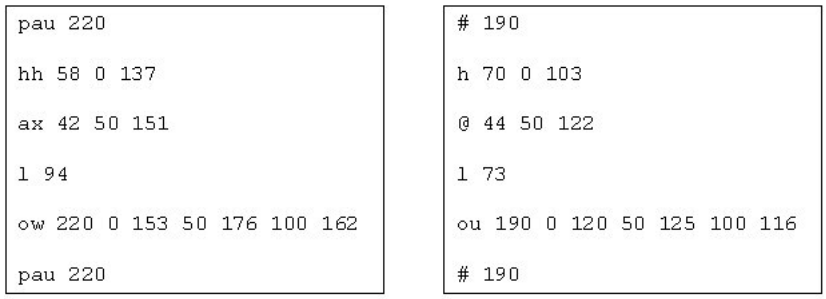
```
(define (Save_Seg_Mbrola_Entry_Ram name start dur targs)
   "(save_seg_mbrola_entry ENTRY NAME START DUR TARGS)
   Entry contains, (name duration num_targs start 1st_targ_pos 1st_targ_val)."
   (set! pho_line (format nil "%s %d " name (nint (* dur 1000))))
   (if targs     ;; if there are any targets
      (mapcar
        (lambda (targ) ;; targ_pos and targ_val
          (let ((targ_pos (car targ))
                (targ_val (car (cdr targ))))
          (set! pho_line
               (string-append pho_line
                (format nil "%d %d "
                 (nint (* 100 (/ (- targ_pos start) dur))) ;; % pos of target

                 (nint (parse-number targ_val)))           ;; target value
           ))
    ))
    targs))
  (set! pho_line (string-append pho_line ""))
)
```

izada para obter o objeto Java que contém a estrutura fonética do texto que acabou de ser sintetizado.

Durante a síntese, existem dois parâmetros que são tratados pelo "Expressive Talking Heads" com o intuito de enriquecer a saída do sistema: o gênero da voz e o idioma da fala. Esses dois elementos são simples de serem manipulados quando já existe uma função que efetua o mapeamento da base de fonemas do Festival para a base de fonemas do MBROLA. A junção desses dois subsistemas trouxe para o "Expressive Talking Heads" o suporte direto ao idioma inglês americano com os gêneros feminino e masculino, e ao idioma inglês britânico com gênero masculino. Para incorporar um novo idioma se faz necessário criar as bases de fonemas para ambos os sintetizadores e estabelecer a função de mapeamento. Existe uma complexidade alta quando uma dessas bases precisa ser definida. Por exemplo, no caso do idioma português, já existe uma base especificada no MBROLA. Porém, o mesmo não é verdade para o Festival. Um trabalho futuro, seria definir a base de dados de fonemas para outros idiomas, em especial o português, no Festival e a respectiva função de mapeamento para as bases já existentes do MBROLA.

A Figura 5.4 ilustra exemplos de informações fonéticas geradas através da síntese da palavra *hello* pelo Festival-MBROLA. Cada fonema é representado pelo seu rótulo, seguido ao lado pela sua duração e os dados de entonação. A entonação é uma lista de zero ou mais pares de valores. Cada par é constituído pela posição relativa do tom (*pitch point*) em termos de percentual da duração do fonema e seu valor correspondente em *Hertz*. Por exemplo, na Figura 5.4 (a) o fonema *ax* possui um tom de 151 *Hertz* exatamente na metade de sua duração (aos 21 milisegundos). A Figura 5.4 (a) é o resultado da síntese utilizando o idioma inglês americano e o gênero feminino, enquanto a Figura 5.4 (b) é o resultado para inglês britânico e masculino. Como pode ser observado, o gênero e o sotaque do idioma influenciam diretamente os fonemas gerados, as durações e as entonações.



Síntese da palavra *hello* sobre a opção *American English Female* (a)

Síntese da palavra *hello* sobre a opção *British English Male* (b)

Figura 5.4: Exemplos de informações fonéticas geradas pelo Festival-MBROLA, na etapa de síntese: (a) idioma inglês americano na voz feminina e (b) idioma inglês britânico na voz masculina.

Como já mencionado, na integração Festival-MBROLA, o MBROLA funciona como o módulo DSP no processo de síntese TtS, sendo responsável por gerar o áudio a partir da estrutura fonética elaborada pelo Festival (módulo NLP). Para que o "Expressive Talking Heads" pudesse ter um maior controle sobre o processo de síntese, foram também desenvolvidas funções em *Scheme* que comandam a construção da saída sonora. A idéia foi permitir que o sintetizador ETHs atuasse como um componente intermediário entre os módulos NLP e DSP. A Função 5.3, denominada *Save_Seg_Mbrola_File* gera um arquivo com a lista de fonemas, a partir dos dados contidos na variável *phoneme_structure* interna ao Festival (a mesma variável

inicializada pela chamada à função *Scheme Synth_Pho_Ram*). Se o sintetizador ETHs tiver interesse em interferir no processo de síntese, antes de chamar a função *Save_Seg_Mbrola_File*, ele cuida de alterar o valor dessa variável com a nova estrutura fonética a ser utilizada, por exemplo, alterando a duração de um determinado fonema.

Função 5.3 Gera um arquivo contendo a estrutura fonética a partir de uma variável interna do Festival denominada *phoneme_structure*.

```
(define (Save_Seg_Mbrola_File)
   (let ((fd (fopen "/tmp/resultado" "w")))
    (fwrite phoneme_structure fd)
    (fclose fd))
)
```

O passo conclusivo na síntese é realizado com a chamada à Função 5.4, denominada *Synth_Wav*. Essa função gera o arquivo de áudio digital contendo a fala sintetizada, a partir de um arquivo de entrada, que no caso, é o arquivo gerado pela função *Save_Seg_Mbrola_File*. O arquivo de som resultante é codificado em formato *wav*.

Função 5.4 Sintetiza o arquivo de entrada, gerando um arquivo de áudio.

```
(define (Synth_Wav)
        "[Synth_Wav: sintetiza o arquivo de entrada gerando um arquivo de áudio]"
        (let ((filename "/tmp/resultado"))
        (system (string-append mbrola_progname " "
                               mbrola_database " "
                               filename " "
                               filename ".wav"))
        )
)
```

Outro elemento no processo de síntese da fala que é levado em consideração no "Expressive Talking Heads" é a pausa. A Seção 5.1.4 destina-se a apresentar as abordagens de tratamento para esse parâmetro que foram estudadas neste trabalho. O principal objetivo dessa investigação foi procurar alterar o comportamento de síntese do Festival, que atribui sempre o mesmo valor para a pausa entre sentenças, a fim de oferecer um resultado mais realista. A investigação do tratamento da pausa também serviu para identificar a forma mais adequada de se determinar quando um texto já sintetizado deveria ser encaminhado para o módulo de sincronização facial.

5.1.4 Tratamento da Pausa

Uma das contribuições do desenvolvimento do "Expressive Talking Heads" foi o estudo de alternativas para o tratamento da pausa entre sentenças do texto sendo sintetizado. Pôde ser observado que o Festival trata apenas as pausas entre sentenças, definindo sempre um mesmo valor, que por sua vez é função do idioma em uso. No entanto, ao observar uma fala natural, é possível perceber que os tempos de pausa sofrem variações, às vezes dependendo inclusive do estado de ânimo do interlocutor.

Esta seção apresenta três abordagens que foram definidas no sistema. Na realidade, o sistema definiu duas possibilidades para o tratamento da pausa. A primeira possibilidade,

que deu origem à primeira abordagem, simplesmente utiliza o tratamento padrão do Festival, onde todas as pausas apresentam o mesmo valor. Já a segunda possibilidade deu origem às outras duas abordagens. Ambas alteram, de maneira aleatória, o tempo para as pausas entre sentenças, interceptando a saída do Festival e editando a descrição fonética antes de enviá-la para geração do arquivo de áudio. As duas abordagens diferem no momento em que a requisição para geração do áudio é feita.

A escolha da abordagem a ser utilizada é feita pelo usuário na interface do sistema. Todas elas iniciam recebendo um bloco de texto proveniente do *parser*, cuja característica é estar inteiramente contido em um mesmo estado de ânimo (conteúdo de um elemento *div*).

Abordagem 1: Bloco-a-Bloco

Na abordagem *bloco a bloco*, a entrada textual fornecida pelo *parser*, que pode conter várias sentenças, é enviada para o sintetizador Festival (módulo NLP) de uma única vez. A descrição fonética gerada como saída pelo Festival (Figura 5.4) é então armazenada na estrutura de dados da animação (Figura 5.1). Essa estrutura irá conter todas as sentenças do texto, sempre com um mesmo valor de pausa entre elas. A mesma estrutura recebida do módulo NLP é entregue pelo Festival ao MBROLA (módulo DSP), que gera o áudio correspondente ao bloco de texto (fala do personagem) inicialmente recebido. Encerrado o processo, o sintetizador ETHs requisita que o módulo de sincronização inicie a animação facial e após seu término limpa a estrutura de dados para a próxima animação. Essa abordagem foi denominada bloco a bloco porque realiza a síntese do texto a cada bloco, tanto para o Festival como para o MBROLA. A Função 5.5 ilustra o pseudo-código utilizado na implementação dessa abordagem.

Função 5.5 Pseudo-código da abordagem *bloco a bloco*.

```
passo 1: envia para o Festival o texto a ser sintetizado
passo 2: chama a função Synth_Pho_Ram para síntese
passo 3: armazena a estrutura fonética gerada pelo Festival
passo 4: requisita, através da função Save_Seg_MBrola_File
  que o Festival salve em arquivo a estrutura fonética a ser utilizada pelo
  MBROLA na síntese
passo 5: requisita, a partir da chamada à função
  Synth_Wav, a síntese da estrutura salva no passo 4,
  para que seja gerado um arquivo contendo o áudio digitalizado
passo 6: inicia a animação facial (lip-sync)
passo 7: limpa a estrutura de dados da animação
```

Os pontos positivos da utilização da abordagem bloco a bloco são a sua simplicidade de implementação e o bom desempenho para blocos de texto não muito extensos. Nessa abordagem a estrutura fonética só precisa ser transmitida uma única vez, do servidor para o cliente. Mais ainda, apesar do algoritmo separar as funções, uma única chamada ao Festival-MBROLA pode resultar na geração da estrutura fonética e do arquivo digitalizado. Na realidade, o processo ocorre dessa forma na implementação. O pseudo-código apresenta as funções de maneira separada apenas para destacar as diferentes etapas do processo.

Por outro lado, esta abordagem diminui o grau de interatividade do usuário com o sistema quando o bloco de texto é muito longo, pois o usuário precisa esperar por uma síntese demorada. Além disso, essa abordagem resulta em tempos de pausa entre sentenças sempre iguais, reduzindo a naturalidade do áudio gerado.

Abordagem 2: Sentença a Sentença

Na abordagem *sentença a sentença*, cada sentença do bloco de texto fornecido pelo *parser* é tratada individualmente. Cada sentença é enviada ao Festival e sua estrutura fonética retornada é encaminhada para uma função no sintetizador ETHs de tratamento do tempo de pausa (processamento intermediário). A nova estrutura fonética é então armazenada na estrutura de dados da animação e enviada de volta ao Festival, que por sua vez a encaminha ao MBROLA, responsável pela geração do áudio correspondente à sentença. Em seguida, o sintetizador ETHs requisita que o módulo de sincronização inicie a animação facial para a sentença e ao final o sintetizador ETHs limpa a estrutura de dados da animação. O processo se repete até que todas as sentenças tenham sido sintetizadas e enunciadas. Essa abordagem foi denominada sentença a sentença porque realiza a síntese do texto a cada sentença, tanto para o Festival como para o MBROLA. A Função 5.6 ilustra o pseudo-código utilizado na implementação dessa abordagem.

Função 5.6 Pseudo-código da abordagem *sentença a sentença.*

```
passo 1: enquanto houver sentença
  1.a  begin
    passo 2: paramSentence recebe a próxima sentença a ser sintetizada
    passo 3: chama a função synthetizeSentence(paramSentence)
        passo 3.1: envia ao Festival texto contido em paramSentence
        passo 3.2: chama a função Synth_Pho_Ram para síntese
        passo 3.3: armazena a estrutura fonética retornada pelo Festival
        passo 3.4: faz o tratamento de pausa, gerando uma nova estrutura fonética
    passo 4: entrega a nova estrutura fonética de volta para o Festival
    passo 5: requisita, através da função Save_Seg_MBrola_File
      que o Festival salve em arquivo a estrutura fonética a ser utilizada pelo
      MBROLA na síntese
    passo 6: requisita, a partir da chamada à função
      Synth_Wav, a síntese da estrutura salva no passo 5, para que seja
      gerado um arquivo contendo o áudio digitalizado
    passo 7: inicia a animação facial (lip-sync)
    passo 8: limpa a estrutura de dados da animação
  1.b  end
```

A primeira vantagem aparente dessa abordagem é o aumento da naturalidade no áudio sintetizado, pelo fato da pausa não ser mais gerada sempre com um mesmo valor. A outra vantagem aparente é a preservação do requisito de interatividade, uma vez que o tempo de espera para uma nova síntese independe do tamanho do bloco de texto. No entanto, na prática essa vantagem ocasiona uma perda de naturalidade e interatividade, já que o tempo despendido, principalmente na geração do áudio (módulo DSP), introduz uma espera que faz com que não haja uma continuidade na fala do bloco de texto. A abordagem acabou por introduzir pausas muito longas entre as sentenças, fazendo com que o comportamento do personagem se parecesse mais com um robô do que com um ser humano.

Abordagem 3: Sentença-Bloco

Como na abordagem anterior, na abordagem sentença-bloco, cada sentença do bloco de texto fornecido pelo *parser* é enviada individualmente ao Festival. Da mesma forma, a saída do Fes-

tival (estrutura fonética) é encaminhada para uma função no sintetizador ETHs de tratamento do tempo de pausa (processamento intermediário). No entanto, a nova estrutura fonética é armazenada na estrutura de dados da animação mas não é enviada de volta ao Festival. Esse processo é repetido para cada sentença do bloco de texto, de forma que a estrutura fonética final armazenada na estrutura de dados da animação é composta pela concatenação das estruturas fonéticas de cada sentença, modificadas pela função de tratamento de pausa.

A estrutura fonética final é então enviada ao Festival para que seja encaminhada ao MBROLA. Desse modo, uma única saída de áudio é gerada para todo o bloco de texto. Em seguida, o sintetizador ETHs requisita que o módulo de sincronização inicie a animação facial para o bloco e ao final o sintetizador ETHs limpa a estrutura de dados da animação. Essa abordagem foi chamada de *sentença-bloco* porque o processamento NLP (Festival) é feito sentença por sentença, enquanto o processamento DSP (MBROLA) é realizado bloco a bloco. A Função 5.7 ilustra o pseudo-código utilizado nesta abordagem.

Função 5.7 Pseudo-código da abordagem *sentença-bloco*.

```
passo 1: enquanto houver sentença
  1.a  begin
    passo 2: paramSentence recebe a próxima sentença a ser sintetizada
    passo 3: chama a função synthetizeSentence(paramSentence)
        passo 3.1: envia ao Festival texto contido em paramSentence
        passo 3.2: chama a função Synth_Pho_Ram para síntese
        passo 3.3: armazena a estrutura fonética retornada pelo Festival
        passo 3.4: faz o tratamento de pausa, gerando uma nova estrutura fonética
        passo 3.5: concatena a estrutura fonética
  1.b  end
passo 4: entrega a nova estrutura fonética de volta para o Festival
passo 5: requisita, através da função Save_Seg_MBrola_File
  que o Festival salve em arquivo a estrutura fonética a ser utilizada pelo
  MBROLA na síntese
passo 6: requisita, a partir da chamada à função
Synth_Wav, a síntese da estrutura salva no passo 5, para que seja
  gerado um arquivo contendo o áudio digitalizado
passo 7: inicia a animação facial (lip-sync)
passo 8: limpa a estrutura de dados da animação
```

Essa abordagem evita o problema dos longos intervalos entre sentenças observado na abordagem anterior, mas por outro lado, pode sofrer com os problemas de interatividade comentados na primeira abordagem, principalmente quando os blocos de texto forem extensos.

As três abordagens propostas para o tratamento do texto de entrada foram implementadas e estão disponíveis para utilização no "Expressive Talking Heads". Embora todas elas apresentem pontos positivos e negativos, na maioria dos casos, a abordagem sentença-bloco é a que oferece o melhor resultado.

Cálculo do Tempo de Pausa

Para as abordagens sentença a sentença e sentença-bloco, o valor do tempo de pausa entre sentenças é calculado de forma idêntica. O novo valor é gerado sobre uma distribuição aleatória uniforme, como ilustra a Função 5.8.

Função 5.8 Pseudo-código do cálculo do tempo de pausa para um fonema.

```
entrada: constante para pausa de sentença, constante para a pausa de bloco
saída: um novo tempo de duração para o fonema de pausa

passo 1: cálculo da nova duração:
                    PAUSA_SENTENCA + (variavel_randomica * PAUSA_BLOCO)
passo 2: retorna a duração calculada
```

A Figura 5.5 exemplifica o funcionamento do tratamento de pausa (processamento intermediário) para uma sentença contendo apenas a palavra da língua inglesa *hello*. Assim como a pausa entre sentenças, as abordagens definidas poderiam ser utilizadas para outros tipos de tratamento, como entonação, duração dos fonemas etc.

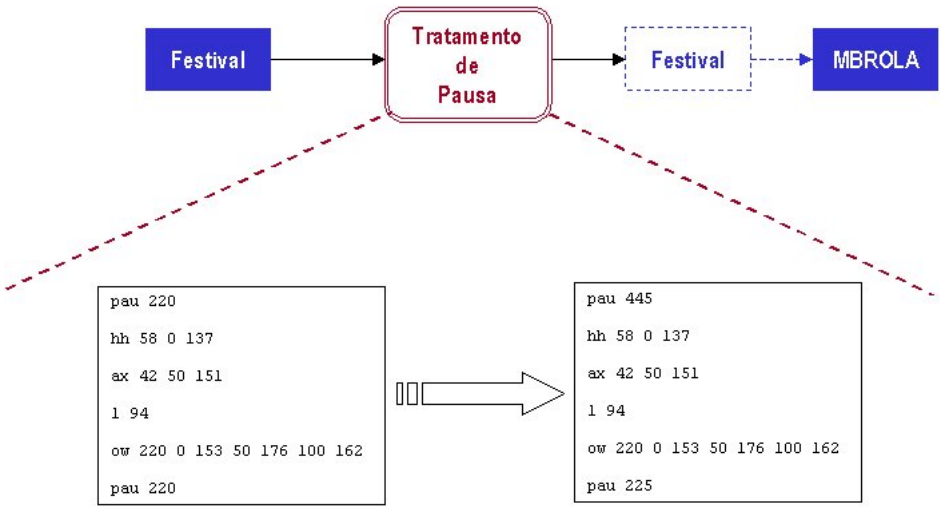


Figura 5.5: Exemplo de funcionamento do tratamento de pausa.

5.2 Módulo de Gerenciamento da Face

A Figura 5.6 ilustra a estruturação do módulo de gerenciamento da face. Esse módulo é responsável por controlar a face do "Expressive Talking Heads" fazendo a ponte de comunicação da interface gráfica do sistema e do módulo de sincronização com o subsistema *Responsive Face*.

Na inicialização do sistema, o módulo de gerenciamento da face carrega as bases estáticas de visemas e expressões faciais, como também uma base que armazena o mapeamento de um fonema para o seu visema equivalente. Durante o processo de animação requisições são feitas pelo módulo de sincronização ao componente de controle da face (*FaceController*) para obter as informações contidas em suas bases. O *FaceController* também recebe requisições para aplicar contrações ou relaxamentos sobre os músculos faciais. Essas requisições são enviadas para a face do sistema controlada por uma instância da classe *Face2bApplet* [Per97]. A interface do sistema também permite que interações do usuário reflitam em transformações nos componentes faciais.

Esta seção destina-se a apresentar a modelagem da estrutura facial no "Expressive Talking Heads". É dada atenção especial à apresentação dos visemas definidos e que são utilizados no momento da animação facial sincronizada com a fala. Por fim, são apresentadas as abordagens assumidas para o tratamento das expressões faciais e dos movimentos de alguns dos

principais componentes faciais (olhos, sobrancelhas e cabeça). A animação da face é descrita na apresentação do módulo de sincronização (Seção 5.3).

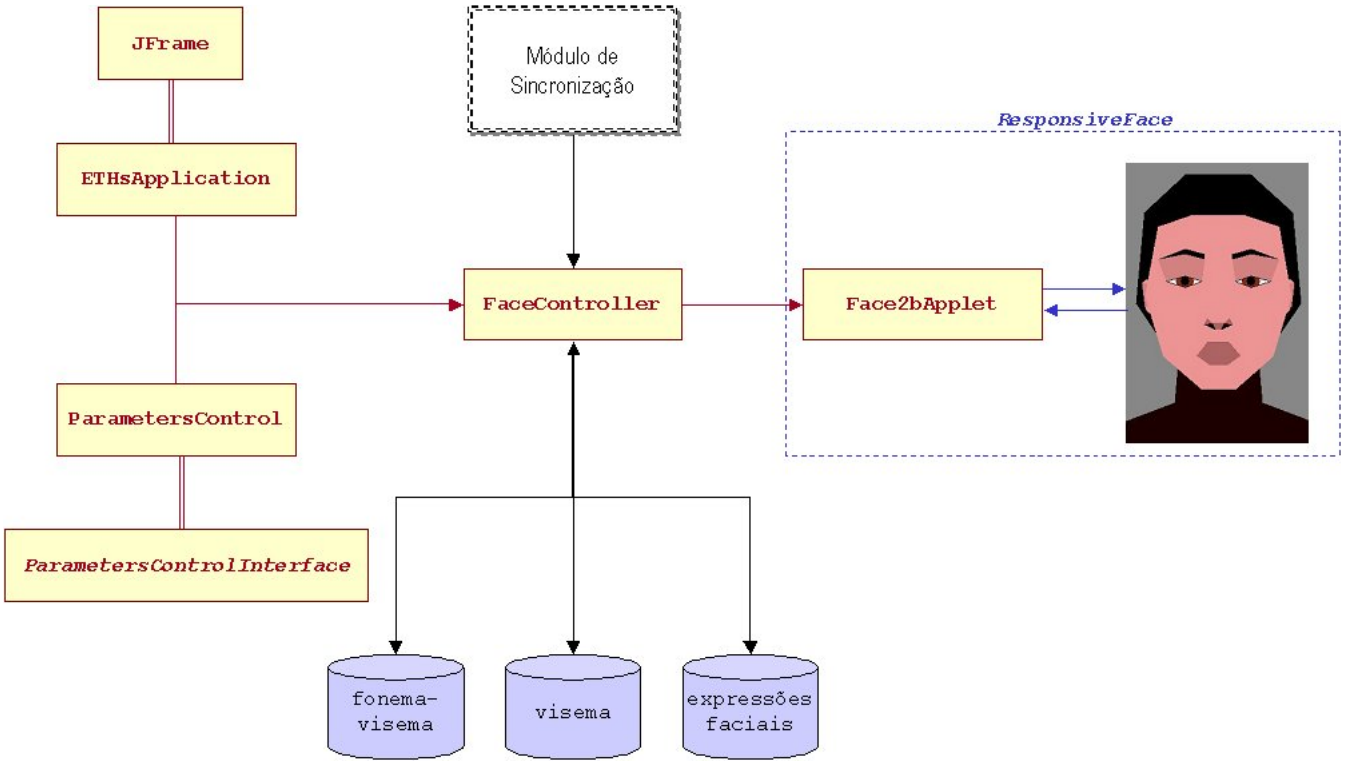


Figura 5.6: Visão geral do módulo de gerenciamento da face.

5.2.1 Unidades Fundamentais

O módulo de gerenciamento da face possui basicamente três unidades fundamentais: as expressões faciais, os visemas e a estrutura fonema-visema.

As expressões faciais correspondem ao estado de ânimo que o personagem virtual pode assumir. Internamente, uma expressão facial (classe *FacialExpression*) é composta por um identificador único e pelos músculos da face que são afetados na expressão (Figura 5.7). Os valores que cada músculo pode assumir estão compreendidos no intervalo fechado [-1.0, +1.0]. As expressões faciais no "Expressive Talking Heads" são definidas através de 12 músculos, agrupados em músculos dos olhos, músculos da boca e músculos do posicionamento da cabeça. A Tabela 5.4 apresenta como exemplo os valores que cada músculo facial assume para a expressão de raiva (*angry*).

A segunda unidade fundamental do módulo é o visema (classe *Viseme*). Conforme definido no Capítulo 2, visema é a representação visual de um fonema. Os visemas no "Expressive Talking Heads" são definidos através de um identificador unívoco e de quatro músculos para o posicionamento dos lábios, *sayAh*, *sayOo*, *smile* e *sneer* (Figura 5.8). Foram estabelecidos 16 visemas para o sistema (Seção 5.2.3), cada um deles sendo representado através de valores fixos para cada um dos quatro músculos. Os visemas foram definidos para a face no seu estado de ânimo natural.

Finalmente, a estrutura fonema-visema (classe *PhonemeViseme*) representa o mapeamento entre fonemas e visemas equivalentes. Esse mapeamento é feito através de uma tabela *hash*, cuja chave é o identificador único do fonema e o valor é o visema associado. A estrutura contém

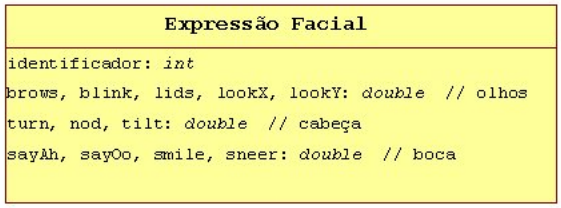


Figura 5.7: A unidade expressão facial.

Tabela 5.4: Valores dos músculos faciais para a expressão de raiva.

Músculo Facial	Valor entre -1 e +1
brows	-1.0
blink	+0.5
lids	+1.0
lookX	0.0
lookY	0.0
turn	0.0
nod	-1.0
tilt	0.0
sayAh	-0.5
sayOo	0.0
smile	-0.7
sneer	+1.0

todos os fonemas possíveis de serem gerados pelo sintetizador TtS, e são mapeados para um dos 16 visemas especificados. Geralmente, um visema representa mais de um fonema (diversos fonemas para um visema). Esta foi a abordagem utilizada no "Expressive Talking Heads", mas seria possível ter para cada fonema um único visema equivalente, ou assumir alguma outra abordagem.

5.2.2 Modelagem da Face

A face no "Expressive Talking Heads" é modelada através de uma malha poligonal tridimensional, herança do projeto *Responsive Face* (Seção 4.2.3 e [Per97]).

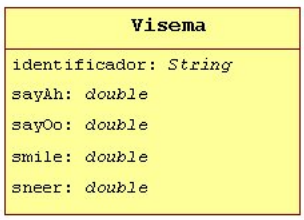


Figura 5.8: A unidade visema.

A estrutura geométrica da malha poligonal definida no *Responsive Face* é ilustrada na Figura 5.9 (b). Com o agrupamento dos vértices da malha, são formados os músculos faciais, sendo os mesmos utilizados para efetuar a animação da face. Dentro do intervalo [-1.0, +1.0], um valor é aplicado a cada músculo para informar o quanto ele vai contrair ou relaxar. Internamente, o *Resposive Face* passa esse valor aos vértices que definem o músculo. A Figura 5.9 (a) exibe a face do sistema com a estrutura poligonal preenchida.

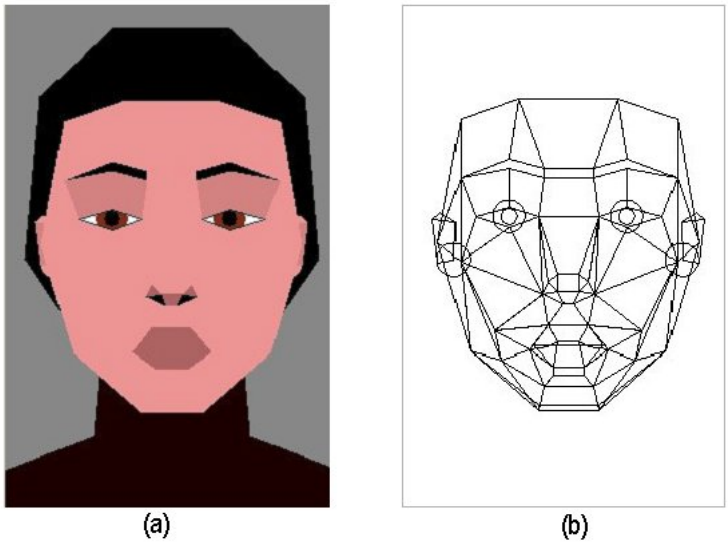


Figura 5.9: Em (a) a face do *Responsive Face* e em (b) sua malha poligonal.

O personagem utilizado nesse trabalho é de estilo caricatural. Como já mencionado, as caricaturas tipicamente envolvem distorcer ou exagerar as características que definem a face em questão. A face no "Expressive Talking Heads" não faz uso de textura, embora seja possível adicionar esse elemento a faces poligonais. Neste trabalho, optou-se por não utilizar o estilo realista (aparência semelhante à face humana), já que o objetivo principal foi alcançar na estrutura facial a naturalidade da fala expressiva com simplicidade e eficiência.

A face do *Responsive Face* possui uma modelagem simples, com mínimos controles, mas que proporciona uma expressividade rica. O "Expressive Talking Heads" acrescenta vida ao modelo poligonal do *Responsive Face*, através da adição da fala e da definição dos visemas, componentes até então não explorados. Mais ainda, o "Expressive Talking Heads" produz uma animação da face enunciando o texto, sincronizando todos esses componentes com aspectos de expressividade.

5.2.3 Visemas

A definição dos visemas neste trabalho tomou por base os visemas especificados no sistema MikeTalk (Seção 3.4.2 e [EP99]), onde foi definido um grupo de 16 visemas para representar as principais posições dos lábios (Figura 3.5) durante o mecanismo da fala.

A Figura 5.10 ilustra o grupo de visemas deste trabalho, considerando o estado de neutralidade para a face do personagem. Para cada visema, a figura também indica os fonemas correspondentes (mapeamento fonema-visema).

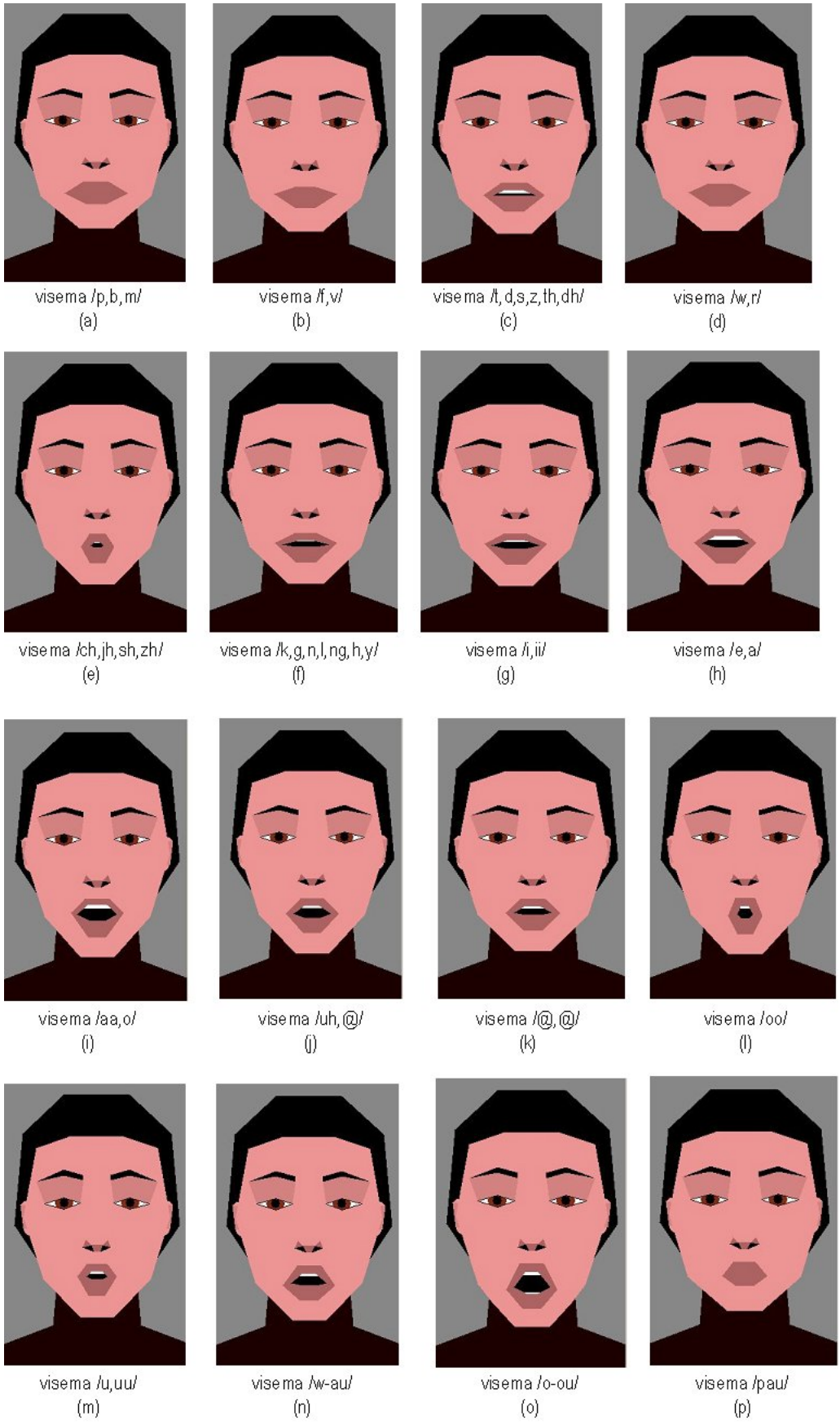


Figura 5.10: Grupo de visemas do "Expressive Talking Heads".

Os visemas do sistema são obtidos através de operações sobre os músculos labiais. Como já mencionado, na face do "Expressive Talking Heads" são definidos quatro músculos desse tipo: *sayAh*, *sayOo*, *smile* e *sneer*. Assim como em toda a estrutura facial, são aplicados valores compreendidos no intervalo [-1.0, +1.0], que definem o quanto cada músculo vai contrair ou relaxar. A Tabela 5.5 apresenta os visemas e suas respectivas posições (valores) para os músculos labiais.

Tabela 5.5: Os 16 visemas do "Expressive Talking Heads" com os respectivos valores para os músculos labiais.

Visema	*sayAh*	*sayOo*	*smile*	*sneer*
/p,b,m/	-1.0	-0.9	0.0	0.0
/f,v/	-0.9	-0.9	0.2	-0.6
/t,d,s,z,th,dh/	-0.5	-0.9	0.24	0.12
/w,r/	-1.0	-0.8	0.28	0.0
/ch,jh,sh,zh/	-0.5	0.62	0.26	-0.4
/k,g,n,l,ng,h,y/	-0.52	-0.74	0.34	-0.36
/i,ii/	-0.4	-0.76	0.48	-0.26
/e,a/	-0.4	-0.64	0.5	0.24
/aa,o/	-0.12	-0.4	0.64	0.24
/uh,@/	-0.36	-0.48	0.5	0.26
/@@/	-0.36	-0.36	0.26	0.24
/oo/	0.0	0.5	0.44	-0.36
/u,uu/	-0.36	0.24	0.6	-0.12
/w-au/	-0.52	-0.72	0.36	-0.26
/o-ou/	0.1	-0.64	0.36	-0.52
/pau/	-1.0	0.0	0.0	0.0

Os visemas são definidos e carregados estaticamente na inicialização do sistema. Eles são armazenados em uma base de visemas, que é continuamente consultada durante a animação facial.

5.2.4 Expressões Faciais

A expressividade é um componente enriquecedor no desenvolvimento de uma animação facial. Quando trabalhada, ela pode ser adicionada a toda a estrutura facial e à fala. No "Expressive Talking Heads" o componente expressividade foi explorado de uma forma preliminar e simplificada, fazendo uso da estrutura já definida no *Responsive Face* (Seção 4.2.3 e [Per97]).

A Figura 5.11 ilustra as expressões faciais que o "Expressive Talking Heads" utiliza. São um total de oito expressões, das quais seis são expressões puras, uma é o estado de neutralidade e uma é uma expressão composta [4]. As expressões puras no "Expressive Talking Heads" são: assustada (*frightened*), que equivale ao medo; desapontada (*disappointed*), que é uma variação da tristeza; incomodada (*annoyed*) que é uma variação do desgosto; surpresa (*surprised*);

[4] As expressões puras são equivalentes às que formam as expressões universais: tristeza, raiva, alegria, medo, desgosto e surpresa, enquanto as expressões compostas são formadas pela composição de duas ou mais expressões puras.

feliz (*happy*); e com raiva (*angry*). A expressão composta é a arrogante (*arrogant*), que é uma composição da raiva com o desgosto.

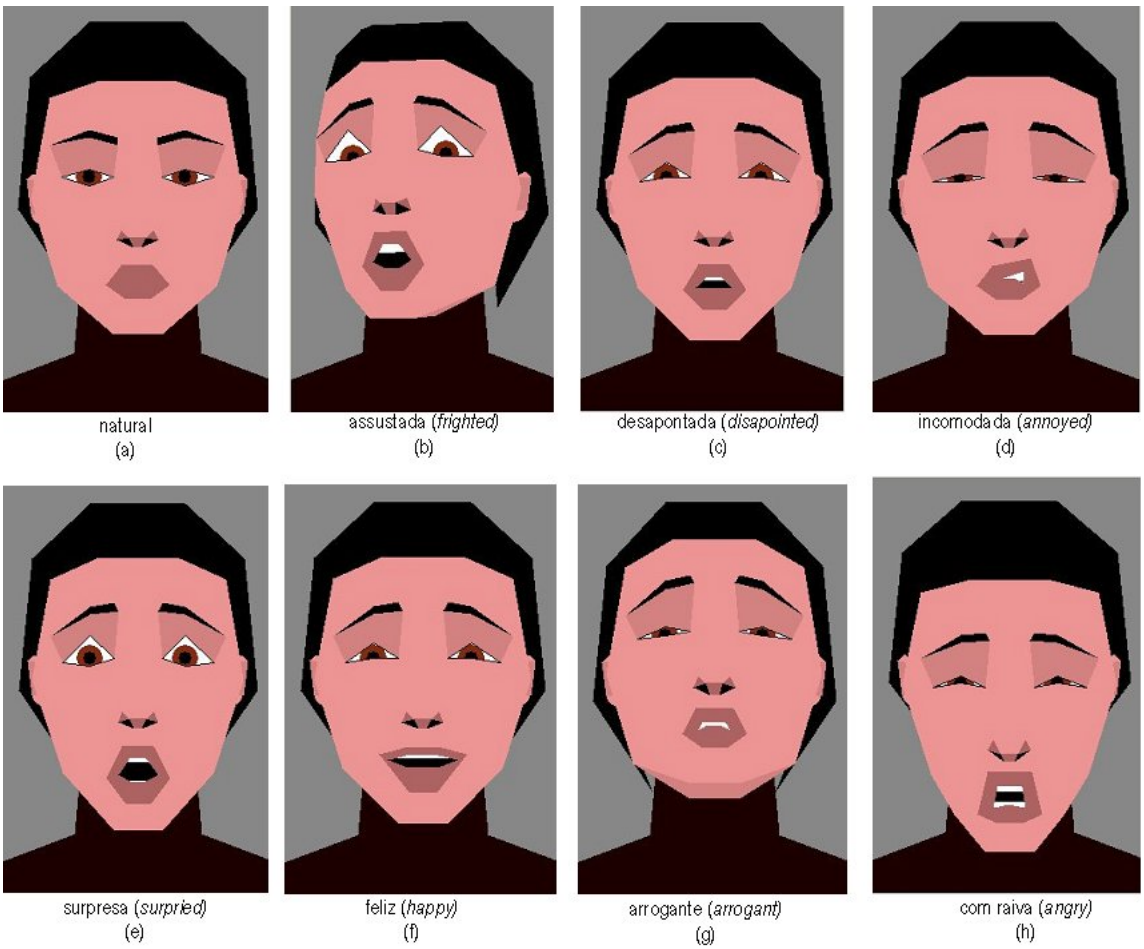


Figura 5.11: Expressões faciais do "Expressive Talking Heads".

A partir das expressões faciais pré-definidas, foram capturados os valores que cada músculo facial deve assumir para cada expressão e, durante a animação, conhecendo-se a emoção desejada, estes valores são aplicados sobre os músculos da face, determinando o quanto cada músculo deve relaxar ou contrair. Como já visto, são utilizados doze músculos faciais para a composição das expressões, dentre os quais cinco correspondem a músculos dos olhos, três a músculos de movimentação da cabeça e quatro a músculos labiais. Cada músculo possui um ponto de atuação e uma função bem definida.

A Tabela 5.6 descreve os valores que cada músculo dos olhos assume para cada expressão facial. Os músculos para o posicionamento dos olhos, pálpebras e sobrancelhas são:

- *brows*: controla o posicionamento das sobrancelhas, que podem estar arqueadas para cima, para baixo ou em um estado de neutralidade (retilíneas).

- *blink*: controla a abertura dos olhos. A princípio são definidos quatro valores para este músculo, proporcionando um olhar "arregalado" (abertura máxima), uma abertura natural (posição neutra), um olhar levemente aberto e o olho fechado (abertura mínima).

- *lids*: controla a direção do olhos, podendo ser para cima, para baixo ou em um estado de neutralidade.

- *lookX*: controla a direção do olhar através das pupilas no eixo de coordenadas X, podendo ser um olhar para a direita, central ou para a esquerda.
- *lookY*: controla a direção do olhar através das pupilas no eixo de coordenadas Y, podendo ser um olhar para cima, central ou para baixo.

Tabela 5.6: As expressões faciais e os músculos dos olhos.

Expressão Facial	*brows*	*blink*	*lids*	*lookX*	*lookY*
natural	0.0	0.0	0.0	0.0	0.0
assustada	+1.0	-0.9	+1.0	0.0	0.0
desapontada	+1.0	0.0	+1.0	0.0	0.0
incomodada	-1.0	-0.5	0.0	0.0	0.0
surpresa	+1.0	-0.9	+1.0	0.0	0.0
feliz	0.0	0.0	+1.0	0.0	0.0
arrogante	-1.0	+0.5	0.0	0.0	+0.5
com raiva	-1.0	+0.5	+1.0	0.0	0.0

De forma análoga aos músculos dos olhos, há três músculos que coordenam os movimentos da cabeça. A Tabela 5.7 descreve o valor que cada um deles assume para as oito emoções do "Expresive Talking Heads". Estes três músculos e suas respectivas áreas de atuação são:

- *turn*: rotação (posicionamento) da cabeça com relação ao eixo de coordenadas X.
- *nod*: rotação (posicionamento) da cabeça com relação ao eixo de coordenadas Y.
- *tilt*: rotação (posicionamento) da cabeça com relação ao eixo de coordenadas Z.

Tabela 5.7: As expressões faciais e os músculos de movimentação da cabeça.

Expressão Facial	*turn*	*nod*	*tilt*
natural	0.0	0.0	0.0
assustada	+1.0	+1.0	0.0
desapontada	0.0	0.0	0.0
incomodada	0.0	0.0	0.0
surpresa	0.0	0.0	0.0
feliz	0.0	0.0	0.0
arrogante	0.0	+1.0	0.0
com raiva	0.0	-1.0	0.0

Por fim, para controle dos músculos da boca nas expressões faciais são definidos quatro músculos. A Tabela 5.8 ilustra o valor dos músculos labiais para cada expressão. Cada um dos quatro músculos atua sobre uma diferente região da boca:

- *sayAh*: abertura vertical da boca.
- *sayOo*: abertura horizontal da boca.

- *smile*: músculo para o sorriso, colocando os vértices das extremidades dos lábios para cima ou para baixo.
- *sneer*: músculo que atua apenas no lábio superior, movimentando-o para cima ou para baixo (boca fechada). O valor deste músculo pode atuar sobre os vértices do nariz.

Tabela 5.8: As expressões faciais e os músculos de movimentação da boca.

Expressão Facial	*sayAh*	*sayOo*	*smile*	*sneer*
natural	-1.0	0.0	0.0	0.0
assustada	0.0	0.0	0.0	0.0
desapontada	-0.5	0.0	0.0	0.0
incomodada	-1.0	0.0	0.0	+1.0
surpresa	0.0	0.0	0.0	0.0
feliz	-0.5	-0.4	+1.0	0.0
arrogante	0.0	0.0	0.0	0.0
com raiva	-0.5	0.0	-0.7	+1.0

5.2.5 Movimento dos Componentes Faciais

Independente do componente expressividade, foi também desenvolvido para o "Expressive Talking Heads" uma função de tratamento de movimento para os músculos da cabeça e dos olhos durante o mecanismo da fala. Estes movimentos podem ser explorados através de cinco abordagens: ausência de movimento, movimento apenas da cabeça, movimento apenas dos olhos, movimento da cabeça e dos olhos sem sincronismo e movimento da cabeça e dos olhos com sincronismo.

As abordagens movimento apenas da cabeça e movimento apenas dos olhos atuam respectivamente sobre os músculos da cabeça e sobre os músculos dos olhos. A abordagem movimento da cabeça e dos olhos sem sincronismo é a composição das duas abordagens anteriores, articulando em conjunto os músculos da cabeça e dos olhos. Por fim, a abordagem movimento da cabeça e dos olhos com sincronismo é uma extensão da sem sincronismo: a diferença é que ela leva em consideração o estado de ânimo do personagem, sincronizando o movimento com essa informação.

Para executar estes movimentos, para cada músculo um valor aleatório no intervalo [0.0, +1.0] é escolhido. Este valor é então multiplicado por n e arredondado, onde n representa o número discreto de posições que o músculo pode assumir. Com isso, é escolhida uma posição aleatória para o músculo dentro de um conjunto discreto de possibilidades. A Função 5.9 descreve o pseudo-código deste tratamento.

A abordagem movimento da cabeça e dos olhos com sincronismo emprega raciocínio diferente para execução. Em primeiro lugar, o cálculo do movimento fica restrito aos músculos que compõem a expressão facial em questão (expressão com a qual o movimento deve estar sincronizado). A outra diferença é que ao invés de calcular valores discretos para os músculos, a função gera um valor dentro do intervalo definido pelo estado de neutralidade e o valor máximo que cada músculo pode assumir na expressão facial correspondente (tabelas 5.6, 5.7 e 5.8) e o aplica diretamente no músculo facial. A idéia é que ao invés do movimento só

Função 5.9 Pseudo-código do tratamento do movimento dos componentes faciais.

```
entrada: MOV_TIME (constante de tempo para mudança de movimento),
FAT_TIME (fator multiplicativo para o cálculo do tempo)
saída:

passo 1: cálculo do tempo de duração de cada movimento:
sleepTime = MOV_TIME + MOV_TIME * (Math.random() * FAT_TIME)
passo 2: escolha da abordagem de movimento dos componentes faciais (fornecido
pelo usuário)
passo 3: cálculo do valor de cada músculo envolvido na abordagem
muscleGroup = (int) Math.round(Math.random() * n_muscleGroup)
passo 4: envia para a face o valor de cada músculo envolvido
```

assumir os valores discretos exportados pelo *Responsive Face*, os movimentos apresentem uma granularidade mais fina.

Independente da abordagem utilizada, o cálculo é efetuado durante toda a fala, propiciando o movimento dos componentes faciais na animação. O efeito destes movimentos equivale a um tratamento de ruído de alta freqüência. Eles favorecem muito o requisito de naturalidade, pois normalmente, quando uma pessoa está falando, ela gesticula, movimenta a cabeça, os olhos, enfim, movimenta o corpo e a face de uma forma natural, espontânea e, muitas vezes, aleatória.

5.3 Módulo de Sincronização

O módulo de sincronização (ou módulo de *lip-sync*) é o elemento mais importante e um dos elementos de maior complexidade no desenvolvimento de um sistema *talking head*, por ser o responsável pela sincronização fina entre a fala e os componentes faciais. Maior ainda se torna essa complexidade se o sistema se propõe a animar a face com interatividade e em tempo real. No "Expressive Talking Heads" o controle de sincronização é feito com o uso de *threads*, que são oferecidas como um mecanismo para programação concorrente na linguagem Java [Lea96] [OW97].

A Figura 5.12 ilustra a visão geral do módulo de sincronização. A operação desse módulo depende da saída gerada pelo módulo de síntese (áudio digitalizado e estrutura de dados da animação) e das unidades fundamentais contidas nas bases do módulo de gerenciamento da face (visemas, expressões faciais e mapeamento fonema-visema). A idéia por trás do funcionamento do módulo de sincronização é, em paralelo com a reprodução do arquivo de áudio digitalizado (sinal da fala), determinar a cada instante o fonema sendo pronunciado e o estado de ânimo do personagem; mapeá-los no visema e nas demais expressões faciais correspondentes, e com isso gerar a animação dos músculos faciais sincronizada com a fala.

O restante desta seção apresenta os detalhes do funcionamento e os relacionamentos dos componentes do módulo de sincronização ilustrados na Figura 5.12.

5.3.1 Unidades Fundamentais

O módulo de sincronização possui basicamente uma única unidade fundamental: a transição de fonemas (classe *PhonemeTransition*). Essa unidade define, para um determinado instante da fala, o fonema sendo pronunciado e o fonema subseqüente. Mais ainda, a estrutura permite

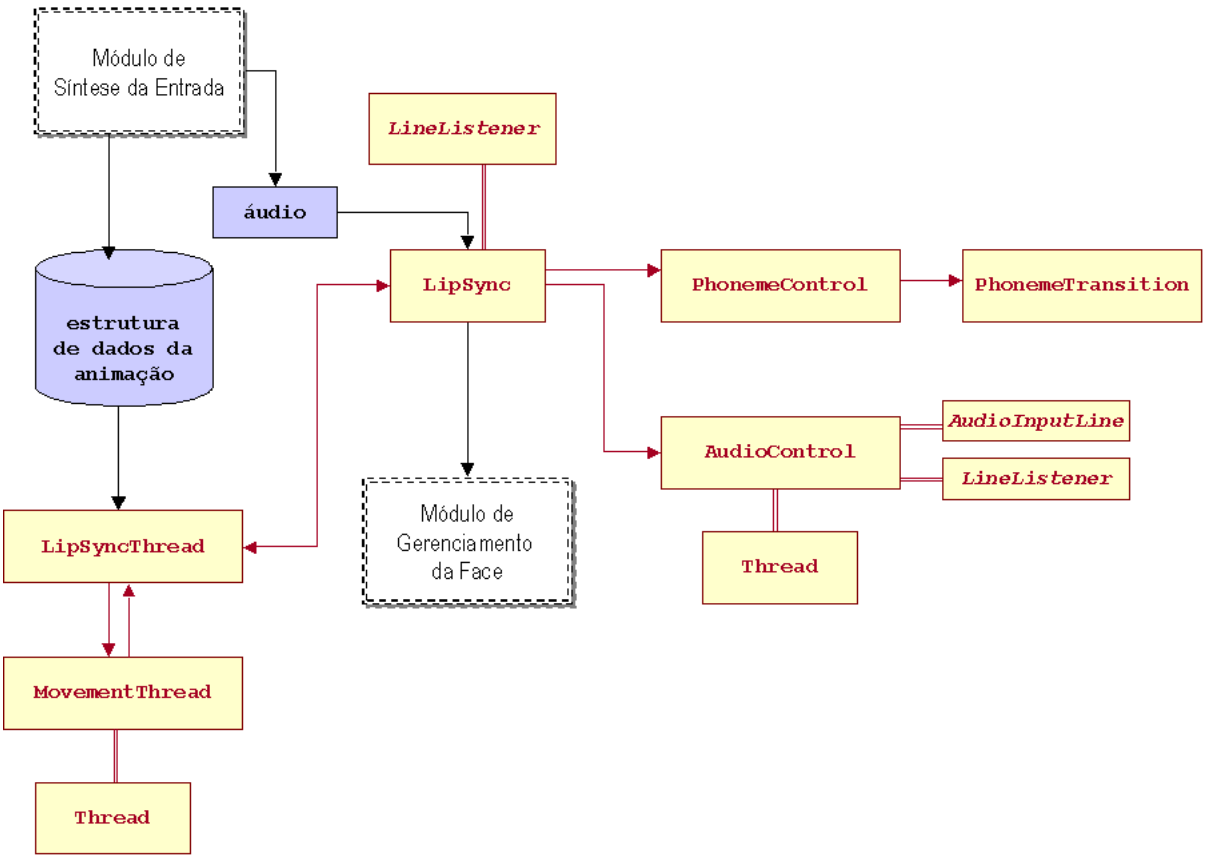


Figura 5.12: Visão geral do módulo de sincronização.

identificar a contribuição de cada fonema em um dado momento da animação. A Figura 5.13 ilustra os atributos da unidade fundamental *transição fonema*.

Figura 5.13: A unidade *transição de fonemas*.

A razão de utilizar uma estrutura contendo dois fonemas foi elaborar a animação facial não apenas com base nas informações dos fonemas contidos na fala, mas também nos aspectos de coarticulação. Desse modo, a animação facial no "Expressive Talking Heads" é feita com base em difones.

O módulo de sincronização utiliza um único objeto *transição-fonema* cujos valores dos atributos variam ao longo da animação. Com base na estrutura de dados proveniente do módulo de síntese, o instante que a reprodução do áudio da fala iniciou e o instante corrente, uma função do módulo de sincronização descobre o fonema sendo apresentado e o fonema seguinte. Utilizando as durações desses dois fonemas, essa função determina, no momento do cálculo, qual o percentual de contribuição para cada um deles e armazena todos esses dados no objeto *transição-fonema*. Esse objeto é então utilizado pelo módulo de sincronização

para definir a transição dos visemas correspondentes aos fonemas. O módulo então requisita ao componente de controle da face que sejam aplicadas as contrações ou relaxamentos nos músculos proporcionalmente à contribuição de cada visema/fonema.

A função desenvolvida considera que um fonema alcança 100% da sua contribuição quando atinge a metade de sua duração. A partir desse ponto, o fonema é considerado o fonema corrente da transição e assim permanece até a metade da duração do fonema seguinte. Nesse instante, a contribuição do fonema deixa de existir, o fonema seguinte passa a ser o corrente, um novo fonema é definido como sendo o seguinte e o processo se repete.

5.3.2 Controle da Fala e da Animação Facial Sincronizadas

O módulo de sincronização labial recebe como entrada a estrutura de dados da animação, basicamente a estrutura fonética da fala, e o arquivo de áudio contendo a fala do personagem. Sua responsabilidade é comandar o módulo gerenciador da face e procurar garantir que a apresentação do áudio esteja sincronizada com a transição dos visemas e com as expressões faciais.

Dessa forma, o primeiro passo na animação é iniciar a reprodução do arquivo de áudio. Pelo mesmo motivo que justificou a escolha da linguagem HTML como estrutura de marcação do texto, o "Expressive Talking Heads" optou por utilizar a tecnologia JavaSound [5] para controle da apresentação do som. O pacote *javax.sound* passou a fazer parte da distribuição da linguagem Java 2 a partir da versão 1.3 e assim desobrigou a instalação de classes externas para execução do sistema. Uma solução alternativa seria o uso do *Java Media Framework* [Mic99], mas foi descartada por obrigar o uso de um pacote externo.

A classe *AudioControl*, ilustrada na Figura 5.12, foi a classe criada para exercer o controle sobre a reprodução do áudio. Ela é iniciada como um *thread* do sistema e instancia um *player* que é um objeto da classe *AudioInputLine* (pacote *javax.sound*). Para tratar os eventos gerados pelo *player*, a classe *AudioControl* implementa a interface *LineListener* (pacote *javax.sound*) e se registra como um ouvinte (*listener*) do *player*. Isso permite que o controlador de áudio do "Expressive Talking Heads" seja notificado no momento exato em que o áudio começa a ser reproduzido. Essa notificação é então passada ao componente de sincronização labial (classe *LipSync*). É muito importante esse tratamento, pois sempre existe um pequeno retardo para início efetivo da apresentação da voz. Como toda lógica da sincronização é relativa ao instante de tempo retornado pelo relógio da máquina no início da animação, é fundamental que esse instante realmente coincida com o início da apresentação do áudio. A implementação da implementação da interface *LineListener* também permite tratar outros eventos do *player* como pausa, retomada, interrupção e o fim natural da exibição.

Em paralelo com a criação e início da execução do *thread* de controle do áudio, o módulo de sincronização (na realidade, o objeto da classe *LipSync*) também dispara um *thread* para o controle da animação facial propriamente dita. A Função 5.10 [6] descreve o funcionamento do *thread* de controle (classe *LipSyncThread*).

O controlador da animação inicia em um laço consultando o objeto da classe *LipSync* esperando pela confirmação do início da apresentação do áudio (Função 5.10 *passo 1*). Essa notificação ocorre da forma explicada anteriormente (classe *AudioControl* e interface *LineListener*).

Quando o controlador da animação recebe a confirmação do início do áudio, é iniciada a

[5] Disponível em http://www.java.sun.com/products/java-media/sound/.

[6] Em alguns pontos, o pseudo-código utiliza a sintaxe da linguagem Java, aproximando-se do código real.

Função 5.10 Pseudo-código do controlador da animação facial.

```
passo 1: enquanto NAO inicio_audio
  1.a  begin
  passo 2: Thread.sleep(5);
  1.b  end
passo 3: init_time = System.currentTimeMillis();
passo 4: movementThread = new MovementThread(getLipSync());
passo 5: movementThread.start();
passo 6: enquanto audio_apresentando
  6.a begin
  passo 7: current_time = System.currentTimeMillis();
  passo 8: phonemeTransition = getLipSync().getPhonemeControl().
    computePhonemeTransition(current_time - getLipSync().
    getAudioControl().getPauseTime() - init_time);
  passo 9: se transicao_fonema igual a NULL
  9.a begin
  passo 10: running = false;
  passo 11: break;
  9.b end
  passo 12: phone1 = phonemeTransition.getFirstPhoneme();
  passo 13: phone2 = phonemeTransition.getSecondPhoneme();
  passo 14: phone1_percent = phonemeTransition.getFirstPhonemePercent();
  passo 15: phone2_percent = phonemeTransition.getSecondPhonemePercent();
  passo 16: viseme1 = getLipSync().getViseme(phone1.getName());
  passo 17: viseme2 = getLipSync().getViseme(phone2.getName());
  passo 18: sayAh = (viseme1.getFlexSayAh() * phone1_percent) +
    (viseme2.getFlexSayAh() * phone2_percent);
  passo 19: sayOo = (viseme1.getFlexSayOO() * phone1_percent) +
    (viseme2.getFlexSayOO() * phone2_percent);
  passo 20: smile = (viseme1.getFlexSmile() * phone1_percent) +
    (viseme2.getFlexSmile() * phone2_percent);
  passo 21: sneer = (viseme1.getFlexSneer() * phone1_percent) +
    (viseme2.getFlexSneer() * phone2_percent);
  passo 22: getLipSync().getFace2bApplet().setLipFlexors(sayAh, sayOo, smile, sneer);
  passo 23: facialExpression.defineFlexorValues(phone1.getFacialExpression());
  passo 24: getLipSync().getFace2bApplet().setEyesFlexors(facialExpression.getBrows(),
    facialExpression.getBlink(), facialExpression.getLids(),
    facialExpression.getLookX(), facialExpression.getLookY());
  passo 25: getLipSync().getFace2bApplet().setHeadFlexors(facialExpression.getTurn(),
    facialExpression.getNod(), facialExpression.getTilt());
  6.b end
```

fase de tratamento dos fonemas. Em primeiro lugar, o controlador consulta a hora corrente (em milisegundos) na máquina em que está executando (Função 5.10 *passo 3*). Em seguida, é instanciado e iniciado um novo *thread* para o controle dos movimentos, conforme explicado na Seção 5.2.5 (Função 5.10 *passos 4 e 5*).

O processo de animação propriamente dito é feito por um laço no *thread* controlador. A cada iteração o controlador faz uma consulta à hora corrente da máquina (Função 5.10 *passo 7*). Essa hora é subtraída do tempo relativo ao início da apresentação do áudio e é então utilizada para descobrir, na estrutura fonética, a transição de fonema correspondente ao instante do áudio (Função 5.10 *passo 8*). De posse da transição de fonema, o sistema obtém para cada um dos dois fonemas que compõe o difone os visemas equivalentes. Essa informação é conseguida numa consulta à estrutura fonema-visema do módulo de gerenciamento da face (Função 5.10 *passos 12 ao 17*).

Os passos 18 a 22 calculam e aplicam as transformações necessárias sobre os músculos labiais para formação dos visemas da face para os fonemas sendo pronunciados. Esses cálculos são feitos com base nas informações de contribuição de cada fonema.

Em paralelo à execução do controlador da animação, ocorre a execução do *thread* de movimento. Ao final da iteração, após aplicar as transformações sobre os músculos labiais, o controlador consulta os valores para os músculos dos demais componentes faciais e aplica as transformações à estrutura da face, buscando oferecer uma fala expressiva (Função 5.10 *passo 23* a *passo 25*). As informações referentes ao estado de ânimo do personagem atualmente são controladas no *thread* de movimento.

5.4 O Expressive Talking Heads como Aplicação *Web*

Como mencionado no decorrer deste documento, o "Expressive Talking Heads" foi projetado para ser um sistema funcionando tanto como uma aplicação de execução local (*stand alone*) quanto como uma aplicação de execução na *web* (*applet*).

A Figura 5.14 ilustra uma visão geral do "Expressive Talking Heads" como aplicação *web*. A estrutura interna de módulos do sistema não foi modificada, apenas foi criado um *applet* (classe *ETHsApplet*) que diretamente se comunica e avisa à aplicação "Expressive Talking Heads" que ela deve executar como uma aplicação *web*.

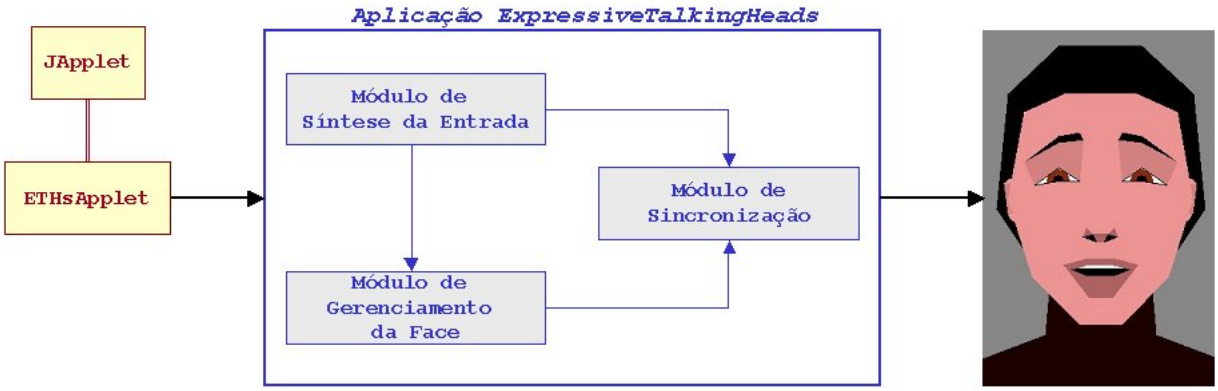


Figura 5.14: Visão geral do sistema como aplicação para a *web*.

Ao iniciar a execução do *ETHsApplet*, alguns parâmetros precisam ser analisados, sendo passados através da página HTML que chama o sistema. Vale ressaltar que estes parâmetros são provenientes da comunicação do sistema com o sintetizador Festival, que é executado

como servidor, como já mencionado, sendo o *ETHsApplet* um cliente dele. O primeiro destes parâmetros é o nome do servidor com o qual o *applet* deve se comunicar. Um segundo parâmetro é a porta de conexão com o Festival-MBROLA. Por fim, o terceiro parâmetro é o endereço em que o sistema deve buscar o arquivo de áudio sintetizado.

Para funcionar como *applet*, as funções de síntese inicialmente definidas para o servidor Festival tiveram que ser refeitas. Quando a única opção de execução do sistema era como aplicação local, o resultado da síntese era apenas armazenado em arquivos: um de fonemas e outro de áudio. As manipulações intermediárias (tratamento da pausa, principalmente) eram realizadas por acessos diretos de leitura e escrita no arquivo de fonemas. Como as restrições de segurança de *applets* impedem que aplicações desse tipo executem qualquer operação sobre arquivos, as funções de síntese foram modificadas de forma a armazenar toda a estrutura fonética em uma variável do sistema (em memória) para apenas no final da síntese ser gerado o arquivo de áudio, conforme foi explicado nas abordagens de tratamento de pausa (funções 5.1, 5.2, 5.3 e 5.4). É importante observar que a gravação final no arquivo de fonemas é feita pelo Festival-MBROLA e não pelo *applet*.

Uma outra restrição de segurança imposta pelo modelo de execução de *applets* obriga que o servidor Festival-MBROLA execute na mesma máquina servidora *web* que hospeda a página HTML e o *applet* do "Expressive Talking Heads". Da mesma forma, o arquivo de áudio utilizado na reprodução da fala, tem que estar localizado nessa máquina e ser acessado através de uma URL do tipo "http://".

Finalmente, uma última alteração foi a eliminação da entrada textual a partir de um arquivo. Na versão *applet* do "Expressive Talking Heads", o texto deve ser digitado de forma interativa pelo usuário. Essa alteração foi feita devido às restrições de segurança impostas aos *applets*: uma delas é a impossibilidade de acessar o sistema de arquivos da máquina cliente.

A rigor, essa restrição de segurança poderia ser removida através da assinatura digital do *applet* e da autorização do cliente que o executasse. Porém, para o "Expressive Talking Heads", essa não é uma boa alternativa pois implicaria na compra de um certificado digital e na conscientização dos usuários de que o programa não irá realizar nenhum tipo de operação indevida em seus computadores.

Capítulo 6

Conclusões

Este trabalho teve como um de seus objetivos estudar os aspectos relacionados às aplicações de animação facial que envolvem a sincronização dos componentes faciais com uma fala sendo pronunciada. Sistemas desse tipo são conhecidos como sistemas *talking head* e são fundamentados em três componentes básicos: a fala, a face e a animação. A partir dos estudos realizados e estabelecendo o requisito de construir um sistema *talking head* com interatividade em tempo real, com tratamento de emoção e buscando oferecer um comportamento natural, próximo à realidade da face humana, esta dissertação resultou no desenvolvimento do sistema batizado de "Expressive Talking Heads".

De forma mais detalhada, o "Expressive Talking Heads" pode ser definido como um sistema que recebe como entrada um texto contendo a fala em conjunto com anotações de expressividade, gênero e idioma, e gera como saída a animação de um personagem virtual enunciando o texto de entrada com o áudio e os componentes faciais sincronizados. O sistema caracteriza-se por explorar a naturalidade da animação facial e ao mesmo tempo por oferecer ao usuário uma interface com interatividade em tempo real. A entrada de dados pode permanecer ocorrendo em paralelo com o processo de animação. O "Expressive Talking Heads" pode executar tanto no modo isolado (*stand alone*) como acoplado a navegadores *web*, na forma de *applet*, tendo sido o sistema projetado e desenvolvido com a preocupação de oferecer uma solução independente de plataforma e sistema operacional. Por essa razão, o sistema foi inteiramente desenvolvido utilizando a linguagem de programação Java (mais especificamente, Java 2) e buscou, sempre que possível, restringir-se ao uso apenas de pacotes incluídos na distribuição básica da linguagem.

O sistema é basicamente composto por um módulo de síntese dos dados de entrada, um módulo de gerenciamento da face e um módulo de sincronização. Cada um desses módulos foca em um dos elementos básicos de um sistema *talking head*, respectivamente: a fala, a face e a animação. Para alcançar as funcionalidades planejadas, o sistema fez uso de três subsistemas já existentes e desenvolvidos em outros projetos: o Festival Speech Synthesis Systems [WT97], o MBROLA [Dea97] e o Responsive Face [Per97]. O Festival e o MBROLA foram integrados ao módulo de síntese da entrada, atuando em conjunto como o sintetizador TtS do sistema. Esse subsistema recebe o texto de entrada previamente analisado por um *parser* do "Expressive Talking Heads" e gera como saída as informações fonéticas do texto e o áudio da fala digitalizado. Já o *Responsive Face* foi integrado ao módulo de gerenciamento da face. Esse subsistema define uma modelagem para a face baseada em uma malha poligonal simples, adequada principalmente para a utilização em navegadores *web*, pois a simplicidade beneficia a eficiência da animação. Apesar de basear-se em uma malha simples, a face oferecida pelo *Responsive Face* é extremamente rica nas expressões faciais geradas. Essa riqueza de expres-

sividade foi explorada pelo "Expressive Talking Heads" para incorporar emoção à animação facial do sistema *talking head*.

Uma outra contribuição deste trabalho foi definir para a face do Responsive Face os valores de contração/relaxamento dos músculos faciais para produzir um conjunto de visemas para serem utilizados na sincronização labial da animação (*lip-sync*). A base de visemas e de expressões faciais para representação das emoções são definidas e tratadas no módulo gerenciador da face.

O desenvolvimento do "Expressive Talking Heads" foi precedido de uma pesquisa sobre os principais elementos que constituem qualquer aplicação desse tipo: a fala, a face e a animação. Para cada um desses elementos foram estabelecidas as abordagens possíveis. Como resultado desse estudo, foi possível propor uma taxonomia para a classificação e comparação dos sistemas *talking head*. A taxonomia proposta engloba os parâmetros fala, face e forma de execução. Para o primeiro parâmetro foram identificadas as abordagens de fala capturada e de fala sintetizada. A fala capturada é resultado da gravação de um áudio, já a fala sintetizada é resultado de uma entrada textual que passa por um sintetizador *Text-to-Speech* [Dut97]. Para o segundo parâmetro foram reconhecidas as abordagens de face definida através de um modelo geométrico e face definida através de imagens capturadas. Em ambas as abordagens, é possível definir uma face bidimensional ou tridimensional, como também definir um estilo para a face em questão podendo ser este caricatural ou realista. Enquanto o primeiro é caracterizado pelo exagero e distorção dos componentes faciais, o segundo caracteriza-se pela sua semelhança com a fisionomia humana. Além disso, para os sistemas construídos a partir de um modelo geométrico, é possível fazer uso de texturas na face, produzindo um efeito de maior detalhe na imagem gerada. Para o terceiro parâmetro (forma de execução) foram consideradas as abordagens de execução em *batch* e em tempo real. No caso dos sistemas *talking head* em *batch* a entrada fornecida pelo usuário é capturada e processada, gerando como saída uma seqüência da animação que será reproduzida quando desejado. Já no caso dos sistemas *talking head* em tempo real, à medida que a entrada é fornecida pelo usuário é gerada a animação facial do personagem virtual enunciando o texto de entrada.

Em particular, o "Expressive Talking Heads" caracteriza-se por ser um sistema de animação facial onde a fala é sintetizada, a face é definida através de um modelo poligonal e sua execução ocorre em tempo real. A execução em tempo real foi muito mais um requisito estabelecido do que uma escolha de abordagem. Esse requisito acabou influenciando nos outros parâmetros: fala sintetizada e modelo geométrico baseado em uma malha poligonal. Com relação à face, é importante salientar a cooperação com a Universidade de Nova York e agradecer ao pesquisador Ken Perlin que ofereceu o código fonte do seu sistema, facilitando em muito o trabalho de integração.

6.1 Contribuições da Dissertação

É possível destacar como resultado desta pesquisa as seguintes contribuições:

- Análise dos elementos básicos que constituem um sistema *talking head*, com a definição de uma taxonomia que permitiu identificar as abordagens possíveis para o desenvolvimento de sistemas desse tipo e com isso estabelecer parâmetros de comparação entre eles.

- Desenvolvimento de um sistema *talking head*, o "Expressive Talking Heads", independente de plataforma e de sistema operacional, podendo executar tanto em modo isolado como acoplado a navegadores *web*.

- Desenvolvimento de um sistema *talking head* com suporte à interatividade e animação em tempo real e tratamento simplificado de emoção na animação de uma face caricatural.
- Integração de subsistemas existentes, Festival, MBROLA e Responsive Face para o trabalho em conjunto no desenvolvimento de um sistema de animação facial com sincronização de fala e expressões. Como conseqüência dessa integração, uma outra contribuição foi a configuração dos sintetizadores Festival e MBROLA para trabalharem cooperativamente. O primeiro funciona como unidade de processamento da linguagem natural (Festival) e o outro funciona como unidade de processamento do sinal digital (MBROLA). A vantagem dessa união foi explorar o maior recurso e configurabilidade do MBROLA para gerenciar bases de voz e idiomas.
- No uso dos sintetizadores de forma integrada, uma outra contribuição do trabalho foi o desenvolvimento de funções em *Scheme* para geração em memória da estrutura fonética e posterior uso dessa estrutura na geração do sinal de áudio digitalizado.
- Definição de três abordagens para o tratamento de pausa, influenciando diretamente na fala do personagem virtual e buscando interferir no processo de síntese para obter um maior realismo.
- Extensão simplificada da linguagem HTML para incorporar parâmetros de idioma, gênero da voz e estado de ânimo, permitindo sintetizar o conteúdo de páginas especificadas nessa linguagem com recursos de expressividade.
- Desenvolvimento de um *parser* para a linguagem de marcação definida que interpreta o texto de entrada e separa as informações de estrutura da fala do conteúdo propriamente dito.
- Definição de diferentes abordagens para tratamento do movimento dos componentes faciais, podendo levar em consideração o estado de ânimo do personagem virtual.
- Construção de uma base de visemas para o modelo facial do Responsive Face, onde cada visema foi definido através de uma combinação de valores para os quatro músculos responsáveis pela expressão labial.

6.2 Trabalhos Futuros

A expressividade é um elemento importante, proporcionando mais vida e naturalidade em uma face. O próximo e principal passo para o trabalho apresentado nesta dissertação consiste em aprofundar a pesquisa sobre os aspectos de expressividade facial, tentando estabelecer métodos e heurísticas para associar a aleatoriedade dos movimentos e contrações musculares com a naturalidade que costuma ser encontrada na maioria das faces.

Um trabalho tão importante quanto a evolução da pesquisa em expressividade é o desenvolvimento de aplicações que utilizam o "Expressive Talking Heads" como base. Na verdade, a utilidade do sistema desenvolvido pode ser melhor comprovada através da construção de aplicações reais. Uma linha de aplicação direta do "Expressive Talking Heads" é o desenvolvimento de *chats 3D* e *conferências virtuais*, onde as pessoas localizadas em diferentes pontos geográficos podem dialogar e trocar informações em ambientes virtuais, utilizando, no caso as funcionalidades de animação e sincronização já oferecidas

pelo "Expressive Talking Heads". No caso da aplicação de *chat*, a face do personagem virtual poderia ser utilizada para enunciar o texto digitado por um usuário remoto. Essa abordagem pode ser uma alternativa interessante a sistemas de vídeo conferência onde a largura de banda oferecida pela rede viabilize apenas a transmissão do áudio, sem o vídeo dos interlocutores. Para esse tipo de aplicação, seria também interessante como trabalho futuro explorar a incorporação de sistemas de reconhecimento de voz para tratar a entrada de dados do usuário. Uma outra linha de aplicações é *shopping virtual online*, podendo o personagem virtual se comportar como um vendedor e interagindo com o cliente, tirando dúvidas e dando sugestões a respeito de uma certa compra. Ainda, uma importante área de aplicação é a de *ensino à distância e treinamento*, onde o personagem pode desempenhar o papel de instrutor, dialogando com o aprendiz e tirando suas dúvidas. Parece ser muito vasta a gama de aplicações que podem explorar os recursos oferecidos pelo "Expressive Talking Heads", sendo as aplicações mencionadas apenas alguns exemplos mais diretos de utilização.

O estado de ânimo do personagem reflete tanto nos componentes faciais como na fala. Atualmente, o parâmetro de expressividade vem sendo trabalhado apenas na face, sendo um interessante trabalho o estudo e desenvolvimento de um novo módulo. Este módulo teria o propósito de incorporar na fala o estado de ânimo de um personagem, alterando o ritmo e a entonação da voz. A adição deste recurso traria um maior realismo à saída do sistema. A princípio, essa incorporação pode ser feita ao sistema atual, sem que seja necessário alterar a estrutura de dados básica atual, a estrutura dos fonemas, pois os mesmos já prevêem os parâmetros de entonação que são inclusive tratados pelo Festival e pelo MBROLA. O trabalho maior seria identificar a maneira adequada de utilizar esses parâmetros.

A utilização de uma linguagem de marcação como forma de entrada dos dados foi também uma importante contribuição deste trabalho, não encontrada em nenhuma das pesquisas relacionadas e discutidas neste documento. A implementação atual baseou-se no padrão HTML, mas com a tendência crescente dos navegadores oferecerem suporte a linguagens baseadas na metalinguagem XML, um outro trabalho futuro seria a utilização de uma linguagem de marcação de propósito específico baseada nesse padrão. Nesse sentido, uma alternativa é investigar o uso da linguagem SSML (*Speech Synthesis Markup Language*) proposta cujo estudo e padronização se encontram em andamento no W3C (*WWW Consortium*). Talvez esse estudo, em conjunto com as pesquisas sobre movimento e expressividade, possa inclusive identificar se a linguagem SSML ofereceria o suporte adequado para as animações faciais que buscam a naturalidade.

Outro tópico de investigação é, a partir do padrão MPEG-4 definido para animações faciais, pesquisar a possibilidade de integração de módulos MPEG-4 similares ao desenvolvido em [Koe01] ao "Expressive Talking Heads". Nessa linha, um outro trabalho seria a implementação de um motor MPEG-4 que utilizasse o *Responsive Face* (e o próprio "Expressive Talking Heads") para realizar animações faciais especificadas em MPEG-4. Isso traria uma nova alternativa para especificar animações no sistema.

A versão atual do "Expressive Talking Heads" trata apenas o idioma inglês (americano e britânico) devido à alta complexidade para a inserção no Festival-MBROLA do suporte ao tratamento de um novo idioma. Uma vez que o MBROLA já possui a base de dados definidas (a versão atual disponibiliza 17 diferentes bases), a dificuldade maior encontra-se no mapeamento do Festival para o MBROLA, visto que o primeiro está restrito a

3 idiomas apenas, sendo necessária a criação do idioma desejado. Fica como trabalho futuro a investigação para definição e integração de novos idiomas ao "Expressive Talking Heads", em particular o idioma português.

Atualmente, a fala da animação é gerada através de um sistema sintetizador, a partir de um texto de entrada. Um outro trabalho de extensão ao sistema seria incorporar um módulo que pudesse capturar a voz de um usuário, permitindo que o sistema trabalhe com as duas abordagens para a fala: capturada e sintetizada.

Por fim, um último trabalho futuro que pode ser mencionado é a personalização dos elementos faciais e da própria fala do sistema *talking head* em função do interesse do usuário. Isto engloba permitir definir a face em termos do gênero e aparência do personagem, aplicar textura no modelo de face e até mesmo identificar as características da voz. Por exemplo, se o usuário do sistema é uma criança, a face do sistema pode se comportar de uma forma diferente, tendo a fala entonações diferentes. Por outro lado, se o sistema estiver sendo usado em uma vídeo-conferência, a fala e a face podem buscar se aproximar o máximo possível das características do próprio usuário.

Apêndice A

Diagramas de Classes do Expressive Talking Heads

Este apêndice destina-se a apresentar os diagramas de classes do "Expressive Talking Heads". Esses diagramas descrevem as classes, suas estruturas internas e seus relacionamentos

Os diagramas de classes estão divididos em: visão geral, síntese da entrada, gerenciamento da face e sincronização.

A Figura A.1 ilustra uma visão geral do sistema "Expressive Talking Heads" através de seu diagrama de classes utilizando a notação UML [RJB99].

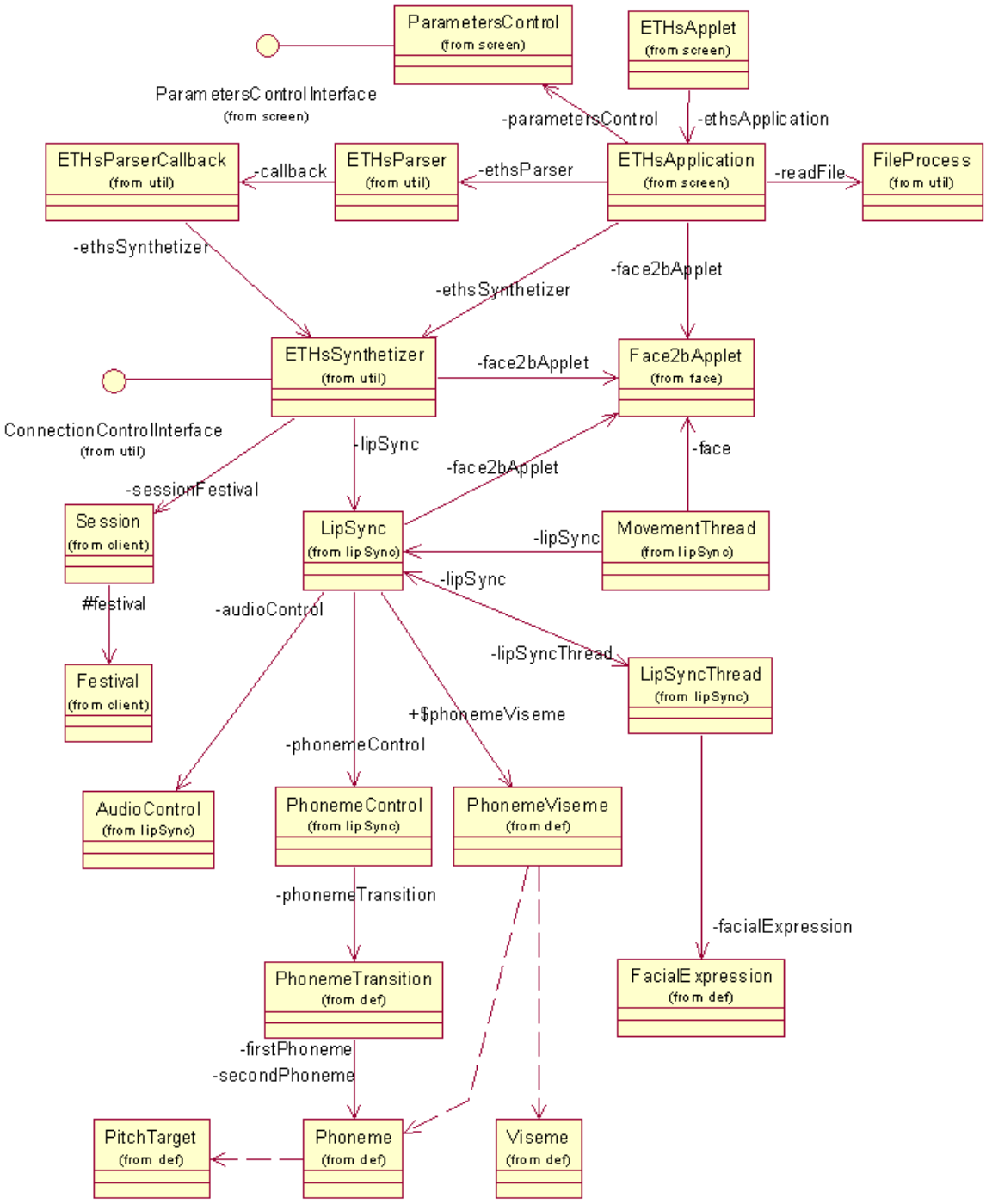


Figura A.1: Visão geral do "Expressive Talking Heads" através de seu digrama de classes.

A.1 Módulo de Síntese da Entrada

O módulo de *síntese da entrada* é responsável por capturar a entrada textual e interpretá-la, separando a fala propriamente dita das anotações sobre a mesma. O texto contendo a fala é enviado para o sintetizador TtS, que retorna como saída as informações fonéticas

do texto e o áudio contendo a fala. A Figura A.2 ilustra o diagrama de classes para este módulo.

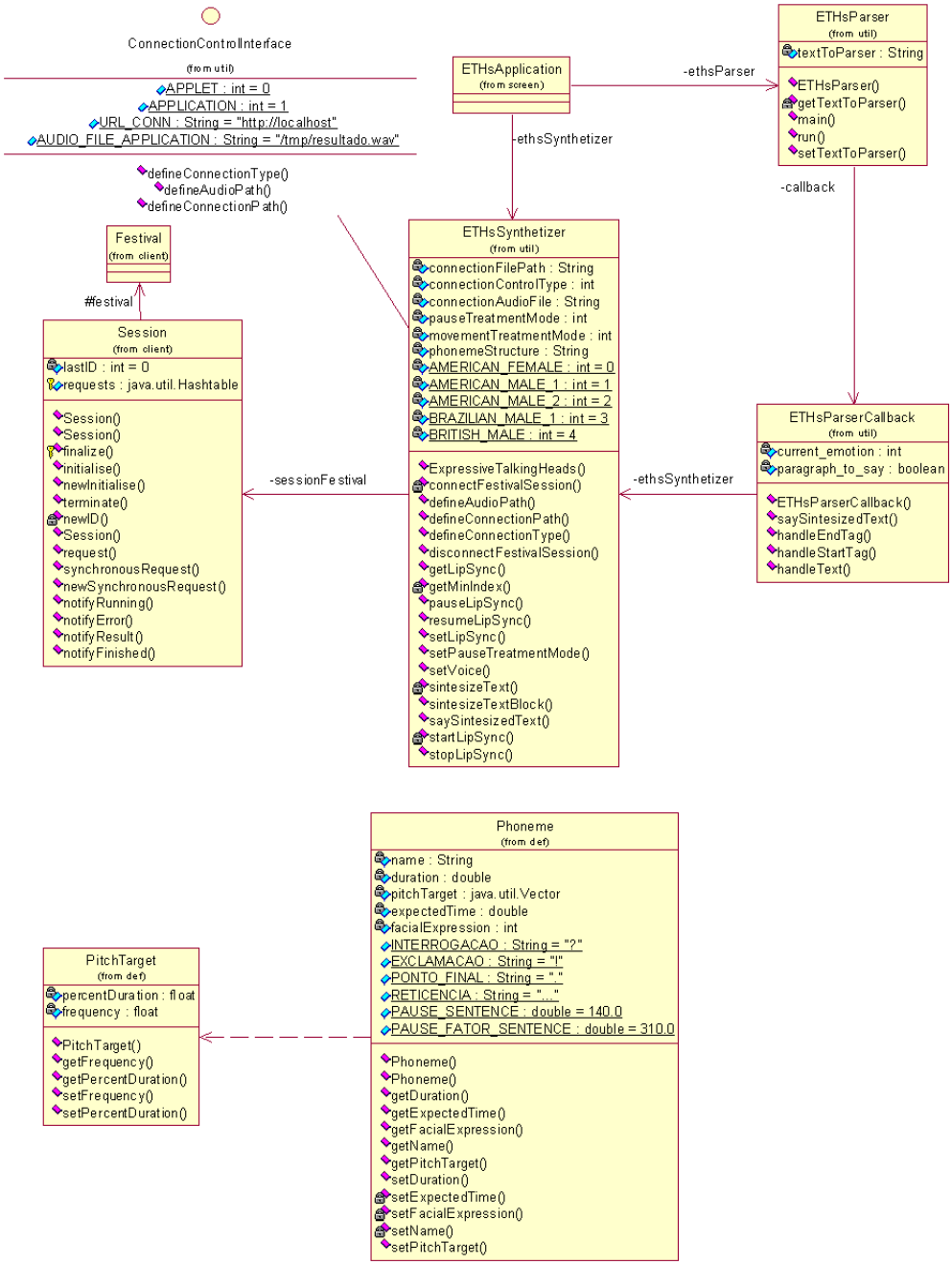


Figura A.2: Diagrama de classes para o módulo de síntese da entrada.

A.2 Módulo de Gerenciamento da Face

O módulo de *gerenciamento da face* é responsável por manter a base de visemas e a base de expressões faciais. A Figura A.3 ilustra o diagrama de classes para este módulo.

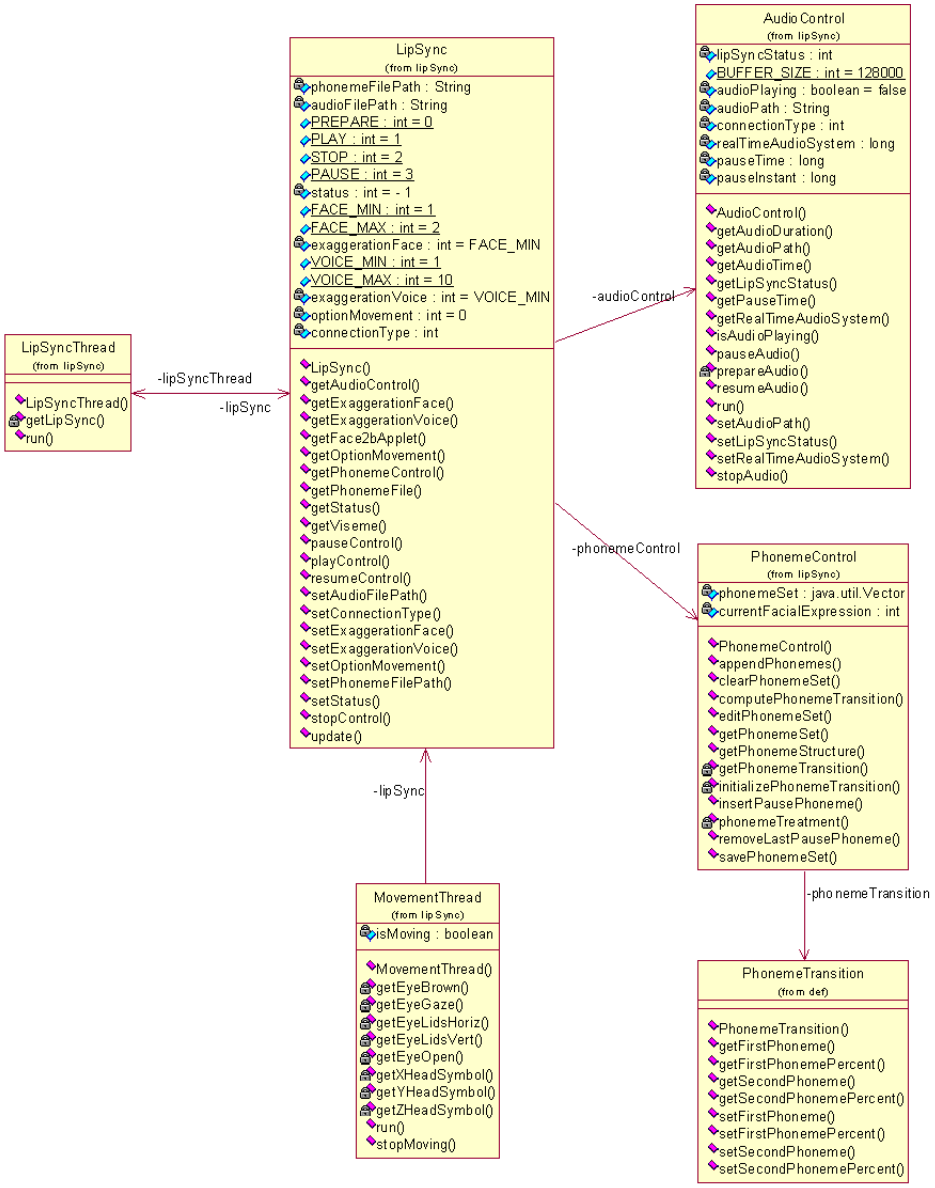


Figura A.3: Diagrama de classes para o módulo de gerenciamento da face.

A.3 Módulo de Sincronização

O módulo de sincronização recebe como entrada a saída proveniente do módulo de síntese da entrada, sendo sua responsabilidade gerar uma animação facial com o conteúdo dos dois outros módulos sincronizados. A Figura A.4 ilustra o diagrama de classes para este módulo.

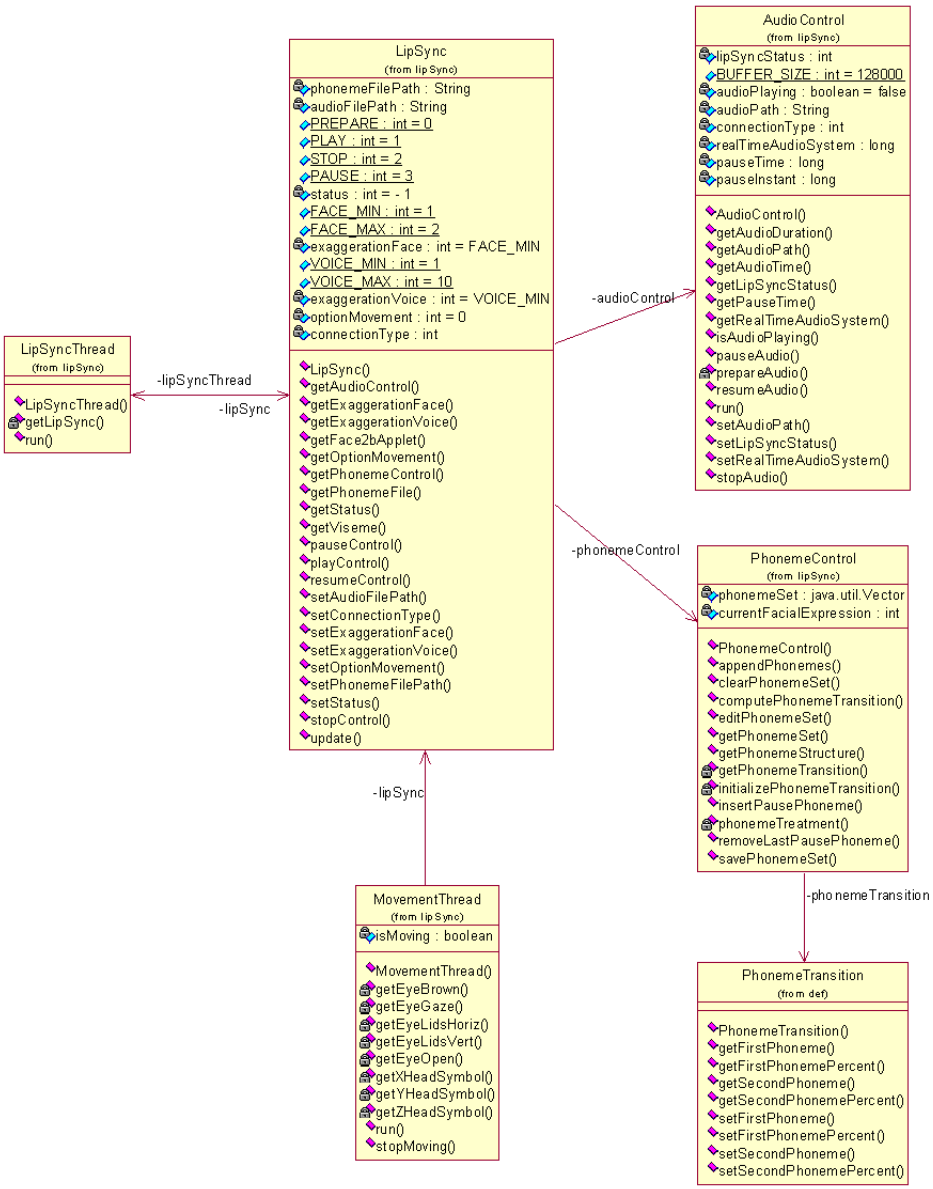


Figura A.4: Diagrama de classes para o módulo de gerenciamento da face.

Referências Bibliográficas

[BCS97] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH '97*, Los Angeles, CA, Agosto 1997.

[Bec01] Evanildo Bechara. *Moderna Gramática Portuguesa*. Editora Lucerna, Rio de Janeiro, 37 edition, 2001.

[Com00] Dirk Van Compernolle. *Fundamentals of Speech Technology*. Katholieke Universiteit Leuven, 2000. Livro eletrônico disponível em `http://www.esat.kuleuven.ac.be/~compi/pub/st_fundamentals/`.

[Dea96] Thierry Dutoit and et al. The mbrola project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *ICSLP*, Bélgica, 1996.

[Dea97] Thierry Dutoit and et al. The mbrola project, 1997. Disponível em `http://tcts.fpms.ac.be/synthesis/mbrola.html`.

[Der98] Olivier Deroo. A short introduction to speech recognition. Technical report, TTS Research Team, TCTS Lab, Faculté Polytechnique de Mons, Bélgica, 1998. Disponível em `http://tcts.fpms.ac.be/asr/introduction.html`.

[DiP01] Steve DiPaola. Facade - stanford facial animation system. Technical report, Stanford University, 2001. Disponível em `http://dipaola.org/stanford/facade/`.

[Dut97] Thierry Dutoit. A short introduction to text-to-speech synthesis. Technical report, TTS Research Team, TCTS Lab, Faculté Polytechnique de Mons, Bélgica, 1997. Disponível em `http://tcts.fpms.ac.be/synthesis/introtts.html`.

[EP98] Tony Ezzat and Tomaso Poggio. Miketalk: A talking facial display based on morphing visemes. In *IEEE Computer Animation*, Philadelphia, CA, 1998.

[EP99] Tony Ezzat and Tomaso Poggio. Visual speech synthesis by morphing visemes. Technical Report 1658, Center for Biological & Computational Learning and the Artificial Intelligence Laboratory da Massachusetts Institute of Technology, Maio 1999. C.B.C.L Papaer No. 173.

[Fac98] Digital faceworks animation software, 1998. `http://interface.digital.com/`.

[GG00] Miguel Grinberg and Alicia Crivicich Grinberg. Magpie. Technical report, Third Wish Software & Animation, Portland, Oregon, USA, 2000. Disponível em `http://www.thirdwishsoftware.com/magpie.html`.

[GMT01] Stephane Gachery and Nadia Magnenat-Thalmann. Designing mpeg-4 facial animation tables for web applications. In *Multimidia Modeling Conference Proceddings*, Los Angeles, CA, Novembro 2001.

[Hea98] Hsiao-Wuen Hon and et al. Automatic generation of synthesis units for trainable text-to-speech systems. Technical report, Microsoft Research, Redmond, Washington, USA, 1998.

[Hos92] John-Paul Hosom. The cslu toolkit: A platform for research and development of spoken-language systems. Technical report, Center for Spoken Language Understanding (CSLU), OGI Campus, Oregon Health Science University (OGI/OHSU), 1992. Disponível em `http://cslu.cse.ogi.edu/toolkit/index.html`.

[Jav95] The java language, 1995. Disponível em `http://java.sun.com`.

[JS] Guy Lewis Steele Jr. and Gerald Jay Sussman. The **Scheme** programming language. Disponível em `http://www.swiss.ai.mit.edu/projects/scheme/`.

[Koe01] Rob Koenen. Overview of the mpeg-4 standard, Março 2001.

[Las87] John Lasseter. Principles of tradicional animation applied to 3d computer animation. In *Computer Graphics, Volume 21,Número 4*, San Rafael, California, Julho 1987.

[Lea86] John Lasseter and et al. Luxor jr., 1986. Curta metragem.

[Lea95] John Lasseter and et al. Toy story, Novembro 1995. Primeiro longa metragem de animação totalmente desenvolvido utilizando ferramentas computacionais. Informações sobre o filme disponíveis em `http://www.pixar.com`.

[Lea96] Doug Lea. *Concurrent Programming in Java: Design Principles and Patterns*. Addison Wesley, Reading, MA, 1996.

[Mic99] Sun Microsystems. Java media framework, v2.0 api specification, 1999. Disponível em `http://java.sun.com/products/java-media/jmf/2.1/specdownload.html`.

[MT89] Nadia Magnenat-Thalmann. Miralab. Technical report, University of Geneva, 1989. Disponível em `http://www.miralab.unige.ch`.

[Ope92] Opengl: The industry's foundation for high performance graphics, 1992. Disponível em `http://www.opengl.org`.

[OW97] Scott Oaks and Henry Wong. *Java Threads*. O'Reilly and Associates, Sebastopol, CA, 1997.

[Pan01] Igor S. Pandzic. Talking virtual characters for internet. Technical report, Porc. ConTEL, Zagreb, Croatia, 2001.

[Pea97] Frederic Parke and et al. Facial animation: Past, present and future panel, 1997. Disponível em `http://mambo.ucsc.edu/psl/sig97/siggraph97-panel.html`.

[Per97] Ken Perlin. Responsive face. Technical report, Media Research Lab, New York University, 1997. Disponível em `http://mrl.nyu.edu/~perlin/demox/Face.html`.

[PW96] Frederic I. Parke and Keith Waters. *Computer Facial Animation*. A K Peters, Ltd., Wellesley, MA, 1996.

[RJB99] J. Rumbaugh, I. Jacobson, and G. Booch. *The Unified Modeling Language Reference Manual*. Addison-Wesley, 1999.

[RVB02] Philip Rubin and Eric Vatikiotis-Bateson. Talking heads. Technical report, Haskins Laboratories, 2002. Disponível em `http://www.haskins.yale.edu/haskins/HEADS/contents.html`.

[Sea01] Steven Spilberg and et al. Shrek, 2001. Informações sobre o filme disponíveis em `http://www.shrek.com`.

[Sho01] Shout3d, 2001. Disponível em `http://www.shout3d.com`.

[SMA94] J.A. Solewicz, J.A. Moraes, and A. Alcaim. Text-to-speech system for brazilian portuguese using a reduced set of synthesis units. In *International Symposium on Speech, Image Processing and Neural Networks*, Hong Kong, Abril 1994.

[SN95] Ralph Steinmetz and Klara Nahrstedt. *Multimedia: Computing, Communications and Applications*. Prentice Hall PTR, Upper Saddle River, NL 07458, 1995.

[W3C99] WWW Consortium W3C. Extensible markup language (xml) 1.0, Dezembro 1999. Disponível em `http://www.w3.org/TR/2000/REC-xml-20001006`.

[W3C00] WWW Consortium W3C. Hypertext markup language (html) 4.01, Outubro 2000. Disponível em `http://www.w3.org/TR/html4/`.

[WT97] Alan Watt and Paul Taylor. The festival speech synthesis system, 1997.

[WTC99a] Alan Watt, Paul Taylor, and Richard Caley. *The Architecture of the Festival Speech Synthesis System*. University of Edinburgh, Março 1999. Disponível em `http://www-2.cs.cmu.edu/~awb/papers/ESCA98_arch/ESCA98_arch.html`.

[WTC99b] Alan Watt, Paul Taylor, and Richard Caley. *The Festival Speech Synthesis System: System Documentation*. University of Edinburgh, 1.4 edition, Junho 1999. Disponível em `http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html`.

[WW92] Alan Watt and Mark Watt. *Advanced Animation and Rendering Techniques - Theory and Practice*. Addison-Wesley, England, 2 edition, 1992.