Expressive Talking Heads: uma ferramenta de animação com fala e expressão facial sincronizadas para o desenvolvimento de aplicações interativas*+

Paula S. Lucena Rodrigues[1], Bruno Feijó[1], Luiz Velho[2]

[1] Departamento de Informática - Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225, Gávea
22453-900 Rio de Janeiro, RJ.

[2] IMPA - Instituto de Matemática Pura e Aplicada
Estrada Dona Castorina, 110
22460-320 Rio de Janeiro, RJ.

pslucena@inf.puc-rio.br, bruno@inf.puc-rio.br, lvelho@impa.br

Abstract. *This paper presents a system called "Expressive Talking Heads", which focuses on facial animation with synchronization between speech and facial movements and expressions. The work integrates several tools, extending their functionalities and, as a result, offering a framework for the development of interactive applications that explore the use of animated faces and synthesized voice. With this research we were able to propose a taxonomy for classifying talking head systems.*

Resumo. *Este artigo apresenta um novo sistema denominado "Expressive Talking Heads" que trata da animação facial tendo a fala sincronizada com os movimentos e expressões da face. O trabalho desenvolvido integra diversas ferramentas, estende suas funcionalidades e, como resultado, oferece uma estrutura para o desenvolvimento de aplicações interativas que explorem o uso de faces animadas e de voz sintetizada. A partir desta pesquisa foi também elaborada uma taxonomia para classificação dos sistemas talking head.*

1. Introdução

A animação facial de personagens tem despertado um grande interesse nos últimos anos. A face humana é interessante e desafiadora simplesmente pela sua familiaridade. Essencialmente, ela é a parte do corpo que é usada para reconhecer indivíduos. 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 normalmente externa os 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 comunicação e interatividade entre os seres humanos. Entre os diversos tipos de sistemas de animação facial, existe uma classe importante de sistemas e que está ligada a este trabalho: são os

* Este trabalho foi realizado com apoio do Fundo Setorial para o Desenvolvimento Tecnológico das Telecomunicações (FUNTTEL), através do contrato 0594/02.

+ A primeira autora agradece ao laboratório Tecgraf/PUC-Rio, pelo suporte tecnológico no desenvolvimento preliminar do sistema, e ao Prof. Luiz Fernando G. Soares, pela valiosa orientação em questões de hipermídia. Os autores também agradecem ao CNPq pelo suporte às suas pesquisas.

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*.

Este artigo tem o propósito de apresentar um sistema *talking head*, que além de sincronizar a fala com movimentos de uma face, busca fazê-lo em tempo real. O trabalho desenvolvido procura atingir uma naturalidade no resultado produzido. Para isso, o sistema combina à animação expressões faciais relacionadas com a emoção do personagem. Também com o intuito de obter uma naturalidade na animação, o sistema introduz algumas estratégias para o tratamento das pausas entre as sentenças pronunciadas e mecanismos simples para controle de movimentos da cabeça do personagem.

O sistema foi batizado como “Expressive Talking Heads” (ETHs) e integra ferramentas para síntese de voz com uma malha simples, porém bastante expressiva, para modelagem geométrica da face humana. O sistema desenvolvido sincroniza o funcionamento desses vários elementos com os recursos mencionados no parágrafo anterior e oferece uma base para o desenvolvimento de aplicações interativas que explorem o uso de voz sintetizada, sincronizada com os movimentos faciais e também com expressões de emoção. Além disso, o sistema permite que tanto o idioma como o gênero da voz sejam parametrizados.

O ETHs espera como entrada um texto marcado, onde as marcações informam justamente os parâmetros de idioma, gênero e também de emoção para serem aplicados à fala e à face de um personagem virtual. Como saída, a ferramenta gera a animação facial em tempo real do personagem, enunciando o texto de entrada com o áudio e os movimentos faciais sincronizados.

O sistema foi desenvolvido buscando servir como base para ser integrado a aplicações que desejem fazer uso de sistemas *talking head* em suas interfaces. Sendo assim, este trabalho também comenta algumas aplicações interativas que foram criadas fazendo uso do “Expressive Talking Heads”. Uma delas oferece uma interface onde o personagem interage com o usuário enunciando o texto digitado. Outra aplicação é uma simples variação da primeira, oferecendo um *chat* entre dois participantes em que o personagem virtual representa o participante parceiro no diálogo. Todas as duas aplicações foram desenvolvidas como aplicações *web* (*applets* Java), sendo que também possuem versões *stand-alone* implementadas.

Este artigo está organizado da seguinte forma. A Seção 2 discorre, resumidamente, sobre os principais elementos que compõem um sistema *talking head*: fala, face e animação; como também sobre uma taxonomia proposta a partir desse estudo para categorização desses sistemas de animação facial. Por fim, esta seção apresenta alguns dos trabalhos relacionados da área e os classifica segundo a taxonomia proposta. A Seção 3 apresenta o sistema “Expressive Talking Heads”, realizando uma análise de cada um dos seus principais módulos. A Seção 4 descreve como utilizar o ETHs no desenvolvimento de aplicações e oferece alguns exemplos implementados. Por fim, a Seção 5 é destinada às conclusões e aos trabalhos futuros.

2. Sistemas *Talking Head*: Elementos Principais, Taxonomia e Trabalhos Relacionados

Esta seção destina-se a apresentar, de forma resumida, alguns conceitos importantes dos principais elementos que compõem um sistema *talking head* e a taxonomia proposta que objetiva analisar as diferentes abordagens existentes para cada parâmetro. Os

parâmetros estudados foram *fala*, *face* e *forma de execução*. Por fim, a seção comenta alguns dos trabalhos relacionados da área encontrados na literatura e faz uso da taxonomia proposta como ferramenta de comparação entre estes sistemas e o ETHs.

2.1. Elementos Principais e Taxonomia

A fala é um importante instrumento na forma de comunicação do ser humano, podendo ser 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. Em um sistema *talking head*, a fala está diretamente relacionada com o áudio que é reproduzido junto com a animação facial, existindo duas abordagens: voz capturada e voz sintetizada [Lu-Ro02]. Na primeira abordagem, voz sintetizada, o áudio é gerado através de um sistema *text-to-speech* (TtS) [Dut97b]. Na segunda abordagem, voz capturada, a fala pode tanto provir de uma pessoa falando em um microfone como de um áudio já capturado e gravado a ser reutilizado.

O segundo parâmetro estudado foi a face. Um conceito importante para o estudo da face e o seu encadeamento com a fala é o de *visema*, que é a representação visual de cada fonema extraído da fala de uma pessoa. Basicamente, uma face pode ser classificada a partir de duas abordagens: se ela é modelada a partir de imagens capturadas ou de um modelo geométrico (imagem sintetizada) [Lu-Ro02]. Em ambas, é possível possuir uma face bidimensional ou tridimensional. Uma vez modelada a face, o passo subseqüente é a sua animação. A animação da face está diretamente relacionada com a forma que a mesma foi modelada. Na face capturada, a animação facial ocorre através da aplicação de técnicas de operações sobre imagens, como a técnica de *morphing*. Já para a face definida através de um modelo geométrico, é comum a utilização de técnicas de animação sobre os músculos faciais [PaWa97]. Por fim, a expressividade é um elemento capaz de enriquecer um sistema *talking head* e foi trabalhada de forma especial no "Expressive Talking Heads", objetivando ter, através da malha poligonal simples utilizada, o máximo de expressividade possível.

O último parâmetro definido para a taxonomia proposta é a forma de execução que um sistema *talking head* pode assumir, existindo duas abordagens, de forma análoga à fala e à face [Lu-Ro02]. A primeira abordagem é a execução em tempo real, que caracteriza-se por ser interativa, tendo a inserção dos dados ocorrendo em paralelo à animação produzida como saída do sistema. A segunda abordagem é a execução *in batch*, que caracteriza-se por ser uma abordagem passiva, onde primeiramente os dados de entrada são capturados e processados para posterior elaboração de um vídeo que será apresentado quando desejado.

2.2. Trabalhos Relacionados

Video Rewrite [BrCS97] é um sistema que faz uso de uma seqüência de vídeo existente para automaticamente criar um novo vídeo com o mesmo contexto porém com uma nova trilha sonora. Aplicações diretas de sistemas como o Video Rewrite são dublagem de filmes, teleconferência e efeitos especiais. Um exemplo 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. É responsabilidade do estágio de análise criar, a partir de uma seqüência original de um vídeo, um banco de dados de exemplos de quadros do vídeo (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 armazenados na base de dados de treinamento. 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. Nessa etapa é necessária a intervenção humana. A coleção de exemplos de imagens rotuladas forma o que é então denominado de modelo de vídeo.

O estágio posterior ao de análise é o estágio de síntese. Nesse estágio, o novo áudio é segmentado e os fonemas obtidos são utilizados para selecionar a seqüência de vídeo contendo os trifones (três fonemas seqüenciais) que mais se aproxima 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 *morphing*. Basicamente, o Video Rewrite combina 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 taxonomia apresentada, 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, pois a saída é construída através de um processamento *in batch*.

Um outro trabalho similar ao "Expressive Talking Heads" é o sistema *MikeTalk* [EzPo99] [EzPo98]. Esse sistema consiste de um sintetizador de fala texto-visual (TTVS - *text-to-visual speech synthesis system*) que procura gerar animações faciais vídeo-realistas. A modelagem da face do personagem busca 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 o Festival [WaTC99a] [WaTC99b], mesmo sistema de síntese de fala utilizado no "Expressive Talking Heads" œomo será mencionado na seção seguinte. O MikeTalk definiu uma base composta por 16 visemas (representação visual do fonema: posição dos lábios, língua etc.) dos quais 6 visemas representam os 24 fonemas consonantais, 7 visemas representam os 12 fonemas monotongos, 2 visemas representam os ditongos e um último representa o visema do silêncio [EzPo99] [EzPo98].

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, e para contemplar este objetivo o MikeTalk faz uso da técnica de *morphing*. Com o áudio sintetizado e a transição fonema-visema definida, cabe ao módulo de sincronização labial fazer com que a animação do personagem e o áudio sejam reproduzidos de forma sincronizada. Um fluxo intermediário de transição de visemas é criado de forma que o módulo de sincronização carrega a transição apropriada de visemas, examinando qual difone (par de fonemas adjacentes) do áudio está sendo apresentado em um dado instante de tempo. A partir dessas informações o sistema constrói um vídeo da animação facial enunciando o texto desejado. Apesar de apresentar uma saída vídeo realista (com base em uma face capturada), o MikeTalk produz uma saída ausente de emoção ficando a animação da face praticamente restrita aos movimentos labiais. Além disso, o sistema é voltado para execuções em *batch*, uma vez que o *morphing* das imagens e a geração do vídeo consomem um tempo de processamento que prejudica a interatividade.

O terceiro trabalho a ser comentado é o *Facade – The Stanford Facial Animation System* [DiP01]. Facade é um sistema de animação facial parametrizado, composto por sete ferramentas (subsistemas) para definição da animação. O Facade faz uso de uma topologia fixa de face poligonal e, através dos 51 parâmetros que definem a face, permite criar tipos faciais, como também os visemas e as expressões faciais. A fala do personagem é introduzida no sistema através de um áudio capturado e para efetuar a sincronização labial dessa fala com a animação facial do personagem o Facade faz uso da ferramenta Magpie [GrCr00]. Essa ferramenta foi associada com uma ferramenta para a fala, o BaldiSync, que é um módulo do CSLU Toolkit [Hos92] para o desenvolvimento de aplicações e pesquisas sobre a fala, responsável pela sincronização da fala arbitrária com movimentos labiais animados. Uma vez construída a fala do personagem, a sincronização com os quadros da animação é feita de forma automática no Facade. Seguindo a taxonomia para sistemas *talking head*, o Facade caracteriza-se por ser um sistema que faz uso de um modelo geométrico para representar a face, a fala baseia-se em um áudio capturado e sua forma de execução é em *batch*, uma vez que é necessário um tempo não desprezível para compilação do vídeo da animação.

O padrão MPEG-4 [ISO02] oferece facilidades para implementação de animações, em particular animações faciais [Oste98], o que favorece o desenvolvimento de aplicações *talking head*. O padrão inclui suporte para animações utilizando tanto áudio capturado como sintetizado, assim como o uso de vídeos capturados ou sintetizados. A referência [GaMa01] descreve a implementação de uma ferramenta de suporte a aplicações *talking head* na *web*, utilizando MPEG-4 como base para a animação das faces. Apesar de oferecer o suporte, o artigo apenas comenta algumas das possíveis aplicações *web* que poderiam fazer uso das animações faciais sincronizadas com áudios, deixando as suas implementações como trabalho futuro.

3. O Expressive Talking Heads

O "Expressive Talking Heads" foi projetado para ser um sistema de animação facial interativo, capaz de interpretar a entrada do usuário[1] e refleti-las na animação final de um personagem virtual. Essas entradas de dados estão intimamente relacionadas com o discurso a ser pronunciado e com a emoção que o personagem deve assumir para cada trecho de fala. Algumas outras características como gênero da voz e idioma da fala também foram previstos para as interações.

Em termos da taxonomia proposta, o primeiro requisito para o sistema foi desenvolvê-lo com uma abordagem que favorecesse a execução em tempo real. Isso direcionou a escolha das abordagens para os outros dois parâmetros. No caso da fala verificou-se que a voz sintetizada seria a abordagem ideal, já que se desejava desenvolver aplicações que oferecessem a possibilidade do usuário interagir continuamente através de uma entrada textual. Para o parâmetro face concluiu-se que a abordagem mais adequada seria a de face sintetizada, em decorrência do uso da malha poligonal favorecer as alterações em tempo real dos posicionamentos dos vértices da malha, tanto para a formação dos visemas durante a fala quanto para as expressões faciais, reforçando a interatividade do sistema.

[1] O usuário pode ser tanto um ser humano interagindo com o sistema, ou um outro sistema compondo a aplicação.

Portanto, é possível definir o ETHs 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. A Figura 1 ilustra uma visão geral do sistema e o restante desta seção destina-se a apresentar, de forma resumida, as principais características e funcionalidades de cada um desses módulos e dos subsistemas acoplados ao "Expressive Talking Heads".

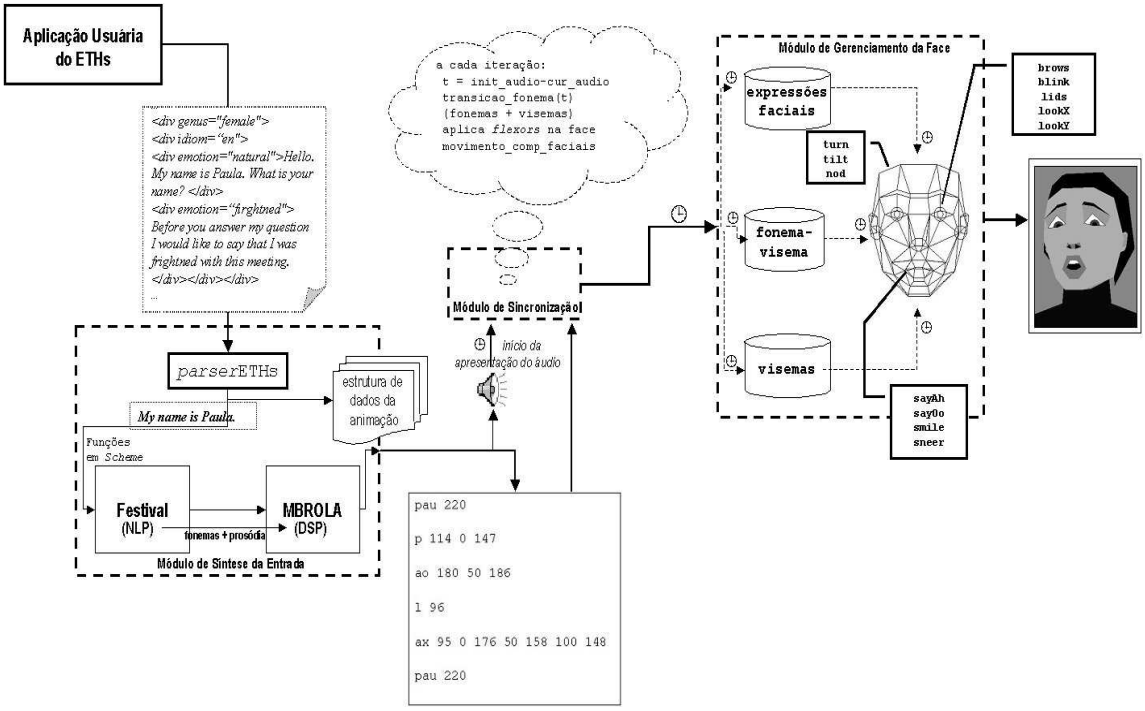

Figura 1 - Visão Geral do "Expressive Talking Heads"

3.1. Módulo de Síntese da Entrada

O *Módulo de Síntese da Entrada* é a parte do ETHs responsável por capturar e tratar o texto fornecido como entrada pelo usuário do sistema e fornecer 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 (exemplificado na Figura 1) 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 britânico). O texto contendo as marcações é capturado e enviado a um elemento denominado *parser ETHs*, que interpreta o texto marcado separando o conteúdo da fala propriamente dita das informações de controle (gênero, idioma e emoção). O fluxo de texto (conteúdo da fala) é enviado para o *sintetizador ETHs* que faz uso dos serviços oferecidos pelos subsistemas Festival [WaTC99a] [WaTC99b] e MBROLA [Dut97a] [Dut96] para obter a descrição fonética e o áudio digitalizado correspondentes à fala do personagem.

No ETHs, os sintetizadores Festival e MBROLA trabalham integrados, de forma que o primeiro atua como unidade NLP (*Natural Language Processing*) e o segundo como unidade DSP (*Digital Signal Processing*). A saída do Festival, que consiste nas

informações fonéticas (os fonemas, onde cada um contém sua respectiva duração e entonação), é entrada para o sintetizador MBROLA, que por sua vez, gera como saída o áudio digitalizado (atualmente formato *wav*), completando assim o processo de síntese. A integração dessas duas ferramentas ao sistema permitiu também que o módulo de síntese do "Expressive Talking Heads" interferisse no processo, buscando colocar um maior realismo na saída de áudio gerada. Nessa etapa, o *sintetizador ETHs* pode ser configurado para acrescentar um tratamento especial à pausa entre as sentenças da fala, interceptando e manipulando a saída fonética do Festival, e repassando essa descrição modificada para o MBROLA a fim de refletir as alterações na saída de áudio gerada [Lu-Ro02]. A comunicação interna entre o Festival e o MBROLA e a integração com o ETHs foram feitas através de funções desenvolvidas na linguagem *Scheme*.

3.2. Módulo de Gerenciamento da Face

O *Módulo de Gerenciamento da Face* é responsável por fazer a ponte de comunicação da interface gráfica do sistema e do módulo de sincronização com o subsistema *ResponsiveFace* [Per97] do qual foi herdada a modelagem da face no ETHs.

A face no sistema é modelada através de uma malha poligonal tridimensional. 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. O ETHs acrescenta vida ao modelo poligonal do *ResponsiveFace*, através da adição da fala e da definição dos visemas, componente até então não explorado. O objetivo principal foi alcançar na estrutura facial já definida a naturalidade da fala expressiva com simplicidade e eficiência.

No módulo de gerenciamento da face tem-se o trabalho co-relacionado de três unidades fundamentais: a estrutura fonema-visema, os visemas e as expressões faciais. Na inicialização do sistema, as bases desses três elementos são carregadas e armazenadas em memória. A estrutura fonema-visema representa o mapeamento entre fonemas e visemas equivalentes. A definição dos visemas no ETHs tomou por base os visemas especificados no sistema MikeTalk (Seção 2.2, [EzPo99] e [EzPo98]), onde foi definido um grupo de 16 visemas para representar as principais posições dos lábios durante o mecanismo da fala. Os visemas do sistema são obtidos através de operações sobre os quatro músculos labiais oferecidos pelo *ResponsiveFace*: *sayAh*, *sayOo*, *smile* e *sneer* (Figura 1). A Figura 2 ilustra o grupo de visemas deste trabalho, como também indica os fonemas correspondentes (mapeamento fonema-visema). Assim como para os visemas, foi definido um grupo com 8 expressões faciais no "Expressive Talking Heads", ilustradas na Figura 3 e aqui citadas: assustada, desapontada, incomodada, surpresa, feliz, com raiva e arrogante. Por fim, a partir das expressões faciais pré-definidas, foram identificados 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.

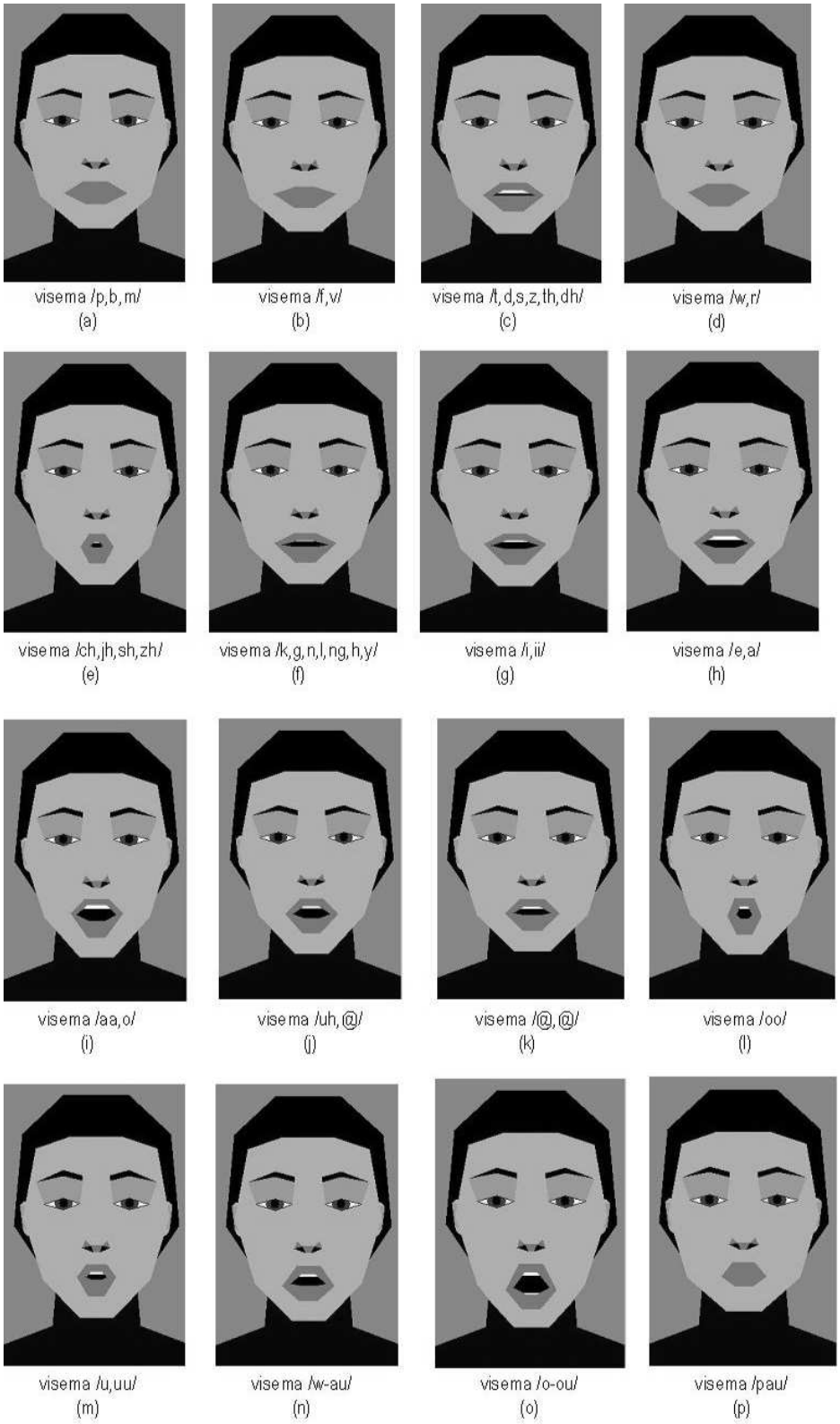

Figura 2 - Grupo de visemas do "Expressive Talking Heads".

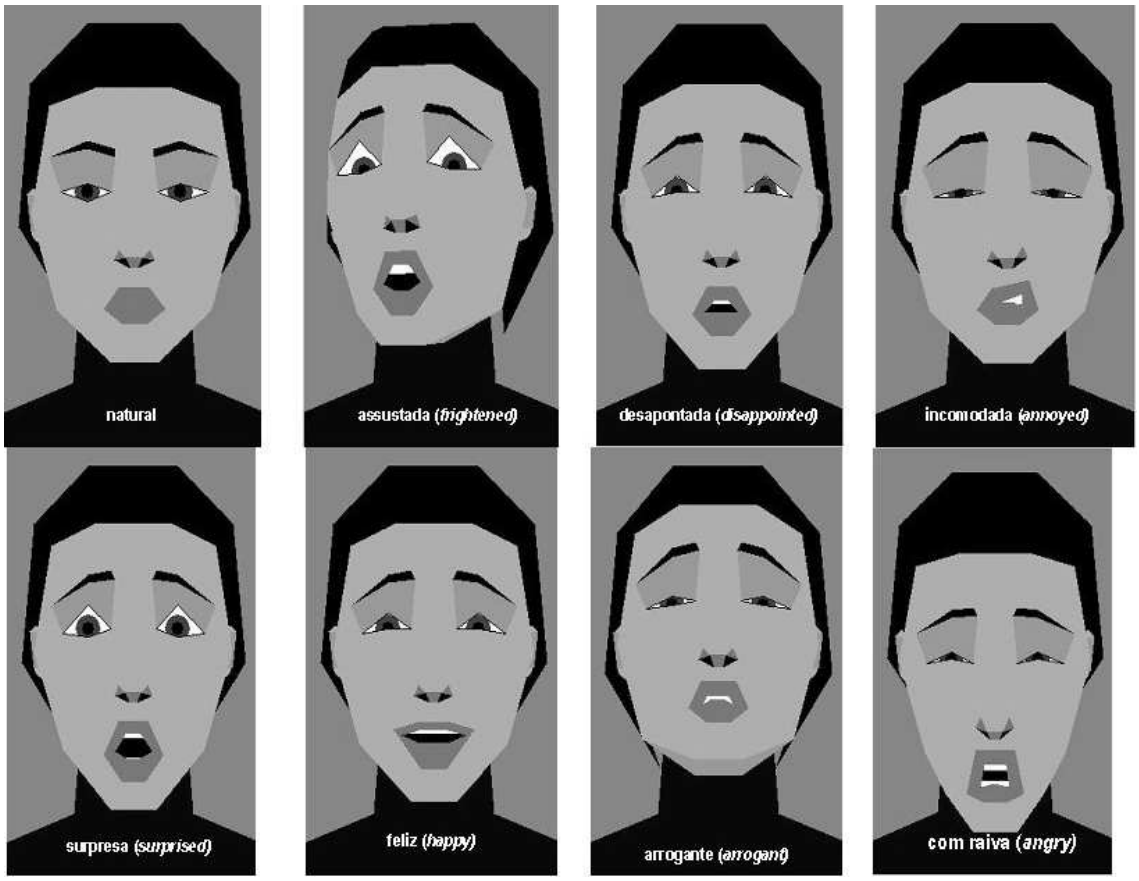

Figura 3 - Expressões Faciais do "Expressive Talking Heads".

3.3. Módulo de Sincronização

O *Módulo de Sincronização* (ou módulo de *lip-sync*) no ETHs 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 propõe animar a face com interatividade e em tempo real.

A operação desse módulo depende da saída gerada pelo módulo de síntese (áudio digitalizado e estrutura fonética da fala) e das unidades fundamentais contidas nas bases do módulo de gerenciamento da face (mapeamento fonema-visema, visemas e expressões faciais). 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. De tempos em tempos uma função do módulo de sincronização descobre o fonema sendo pronunciado e o fonema seguinte. Utilizando as durações desses dois fonemas (obtida do módulo de síntese), essa função determina, no momento do cálculo, qual o percentual de contribuição para cada um deles e armazena todos esses dados em um objeto denominado *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. Dessa forma, o sistema obtém dinamicamente uma representação visual para os difones. Difones são pequenas seqüências de áudio, amostradas através da

transição de um fonema para o meio do fonema subseqüente. A razão para utilizar difones advém da importância de capturar os aspectos de co-articulação da fala. Muitas vezes, a posição dos lábios para um mesmo fonema muda de acordo com o contexto em que o fonema se encontra inserido. De maneira similar, o módulo de sincronização identifica o estado de ânimo corrente a cada verificação e solicita ao módulo de gerenciamento da face que também aplique as transformações necessárias nos músculos da face.

O módulo de sincronização também possui componentes para controle dos movimentos da cabeça e dos olhos. O sistema permite que qualquer um dos dois controles seja ligado ou desligado de forma independente. Quando acionados, esses componentes procuram acrescentar uma aleatoriedade na expressão a fim de aproximar a saída de um resultado mais natural.

4. Aplicações baseadas no Expressive Talking Heads

O "Expressive Talking Heads" foi concebido com o propósito de ser uma ferramenta de apoio ao desenvolvimento de aplicações interativas de animação facial com a fala e as expressões faciais sincronizadas. Esta seção objetiva apresentar a metodologia empregada na construção de uma arquitetura genérica para sistemas *talking heads*, exemplificando como o ETHs pode ser integrado a novas aplicações.

4.1 Metodologia Empregada

O sistema ETHs foi desenvolvido seguindo o paradigma de orientação a objetos e implementado utilizando a linguagem de programação Java.

Como apresentado na seção anterior, o ETHs é composto pelos módulos de síntese da entrada (classe `ETHsSynthesisController`), gerenciamento da face (classe `ETHsFaceController`) e sincronização (classe `ETHsLipSyncController`). Para facilitar o uso do sistema, foi definida uma fachada (classe `ETHsFacade`) [Gamm95], que internamente instancia objetos para os diversos módulos e oferece para as aplicações usuárias uma interface única de acesso, conforme ilustrado no diagrama da Figura 4. A maioria dos métodos oferecidos na fachada possuem um método correspondente em uma das classes internas, onde normalmente a função é efetivamente desempenhada.

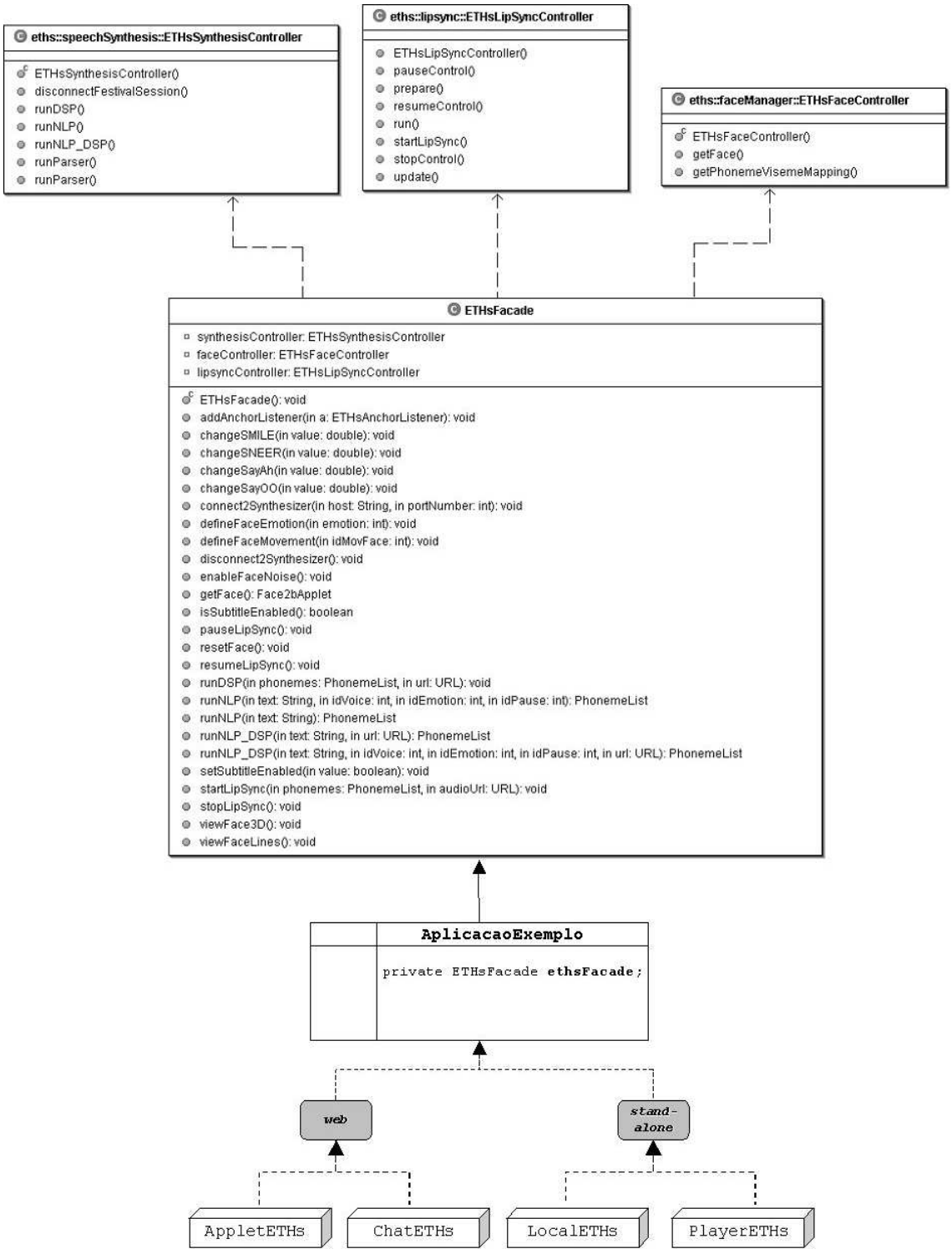

Figura 4 – Diagrama das principais classes do "Expressive Talking Heads".

Existem duas maneiras de uma aplicação usuária do ETHs solicitar a síntese de um texto. A primeira é através da chamada seqüencial dos métodos `runNLP` e `runDSP`. O método `runNLP` espera um objeto do tipo `String` contendo o texto a ser sintetizado, a

identificação da voz[2] (gênero + idioma) e a identificação da emoção, e retorna uma lista de fonemas, onde cada fonema possui sua identificação, duração, entonação e a emoção desejada. Uma vez terminada a execução do método `runNLP`, o método `runDSP` deve receber a estrutura equivalente à retornada pelo processamento da linguagem natural (NLP – Figura 1) e a URL de um recurso onde o ETHs tenha permissão de escrita. Após a execução do método `runDSP`, a URL passada como parâmetro irá conter o arquivo com a fala digitalizada. O oferecimento desses dois métodos de forma separada, permite que uma aplicação interfira no processo de síntese, por exemplo modificando a entonação e duração dos fonemas antes que o áudio seja digitalizado. É permitido que as pausas entre sentenças e entre parágrafos sejam alteradas segundo um modelo probabilístico, objetivando ter uma fala mais natural.

Para aquelas aplicações que não precisam atuar nesse nível, o sistema oferece o método `runNLP_DSP`, que na prática executa seqüencialmente os dois métodos anteriormente comentados. Para o método `runNLP_DSP`, o ETHs permite que seja também escolhida uma estratégia para o tratamento de pausa, sendo que essa estratégia possibilita apenas um cálculo aleatório para a pausa entre parágrafos, não sendo possível alterar a pausa entre sentenças como é feito no processo seqüencial `runNLP` e `runDSP`.

Com a etapa de síntese concluída, a aplicação usuária está habilitada a requisitar que a animação facial, com a sincronização dos lábios e das expressões faciais, seja efetivamente iniciada. Isso é feito com a chamada ao método `startLipSync`.

A chamada ao método `startLipSync` pode receber ainda uma estratégia para o controle dos movimentos dos olhos e da cabeça. Essas chamadas são sempre assíncronas, iniciando uma *thread* (instância da classe `ETHsLipSyncController`) para o controle da sincronização. A animação facial pode ser interrompida, pausada ou retomada através de chamadas aos métodos `stopLipSync`, `pauseLipSync` ou `resumeLipSync`, respectivamente.

Por fim, o método `getFace` retorna uma referência para o componente gráfico contendo a face para que a aplicação tenha controle sobre o posicionamento do rosto do personagem na interface com o usuário.

4.2 Exemplos de Aplicações Desenvolvidas

Como mencionado anteriormente, a ferramenta aqui descrita vem sendo utilizada para o desenvolvimento de diversas aplicações. A primeira aplicação desenvolvida foi batizada com o próprio nome da ferramenta, "Expressive Talking Heads", e possui a característica de executar como aplicação local (*stand-alone*) e de permitir que todos os possíveis parâmetros que a ferramenta oferece sejam habilitados, tendo o máximo de flexibilidade. Essa primeira aplicação é denominada na Figura 4 como *LocalETHs* e permite que o texto fornecido como entrada seja proveniente da leitura de um arquivo ou da digitação do usuário na interface gráfica.

Uma outra aplicação já desenvolvida utilizando a estrutura do "Expressive Talking Heads" é o *AppletETHs*, que oferece funcionalidade semelhante a da aplicação anterior, porém executando em um navegador *web*.

---

[2] A classe `ETHsFacade` define também constantes para as vozes, emoções e estratégias disponíveis. No entanto, essas constantes foram omitidas para evitar uma sobrecarga na figura.

Para implementação dessa aplicação, as classes do "Expressive Talking Heads" foram colocadas em um servidor *web*, juntamente com o sintetizador Festival (operando no modo servidor). Quando executado no navegador, para cada bloco de texto digitado pelo usuário, o *AppletETHs* solicita a realização da animação completa à sua instância do `ETHsFacade`, que por sua vez requisita ao `ETHsSynthesisController` que seja gerado o áudio sintetizado. Nesse momento, o `ETHsSynthesisController` precisa solicitar que o Festival-MBROLA (executando no servidor WWW de onde o applet originou) realize o processamento de síntese da fala. Atualmente, esse pedido é feito através de uma comunicação utilizando *sockets* e seguindo o protocolo particular oferecido pelo Festival-MBROLA. Como trabalho futuro, pretende-se implementar um *proxy* para comunicação com o Festival-MBROLA que ofereça o serviço de síntese seguindo a proposta de *Web Services* [W3C03]. Isso eliminará, principalmente, a necessidade de oferecer serviços externos em uma porta TCP, que normalmente é filtrada nos firewalls.

Quando completada a síntese, o Festival-MBROLA disponibiliza o arquivo no servidor WWW, que é obtido pelo *AppletETHs* através do protocolo HTTP. Além de obrigar que o sintetizador Festival-MBROLA execute na mesma máquina servidora *web*, que hospeda a página HTML e o *applet* do "Expressive Talking Heads", na versão *web* do "Expressive Talking Heads", o texto deve ser sempre digitado de forma interativa pelo usuário, não sendo possível ter acesso ao sistema de arquivos da máquina cliente.

Uma terceira aplicação que foi desenvolvida é o *ChatETHs*, que é uma pequena variação do *AppletETHs*. A idéia por trás dessa aplicação é utilizar a face do personagem virtual para enunciar o texto digitado por um usuário remoto. O sistema oferece uma interface no navegador *web* para escolha do par no diálogo e inicia um processo de cruzamento, onde os applets, ao invés de buscarem o áudio digitalizado do próprio texto, requisitam o áudio do usuário par no *chat*. Através de um protocolo simples de *polling*, o *ChatETHs* consulta o servidor regularmente para saber da existência de uma nova fala a ser enunciada.

Atualmente, uma quarta aplicação vem sendo desenvolvida utilizando o "Expressive Talking Heads". Denominada na Figura 4 de *PlayerETHs_Hyperprop*, essa implementação procura fazer a ponte do sistema *talking head* com um formatador hipermídia [SoMR00]. O objetivo é explorar os recursos da animação facial na elaboração de apresentações hipermídia.

5. Conclusões

Animação facial é uma área importante na Computação Gráfica, cujos avanços de pesquisa influenciam diretamente várias aplicações, com destaque para jogos e filmes/desenhos de animação 3D. Quando integrados a aplicações *web*, personagens *talking face* (ou *talking head*) podem ser de grande utilidade [GaMa01]. Exemplos de aplicações que podem se beneficiar desse recurso são ensino a distância, comércio eletrônico, jogos interativos, *chats* 3D etc.

Este trabalho apresenta como principal contribuição o desenvolvimento de uma ferramenta de animação facial com fala e expressões sincronizadas, fornecendo apoio para a construção de aplicações que desejem fazer uso de interfaces *talking head*. O sistema foi desenvolvido buscando oferecer animações em tempo real, com independência de plataforma e facilidade de interatividade. Como conseqüência, a ferramenta tornou-se adequada para incorporação em aplicações *web*.

Alguns trabalhos futuros podem ser destacados no que diz respeito à expressividade das faces animadas. A expressividade é um elemento importante, proporcionando mais vida e naturalidade em uma animação facial. Sendo assim, uma continuidade do trabalho 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. Atualmente, o parâmetro de expressividade vem sendo trabalhado apenas na face, sendo também um interessante trabalho a inclusão de um componente no módulo de síntese que teria o propósito de incorporar na fala o estado de ânimo de um personagem, alterando o ritmo e a entonação da voz. O uso em conjunto dos sintetizadores Festival e MBROLA beneficia o tratamento da entonação, que inclusive já está contida na estrutura de cada fonema. O maior desafio é identificar a maneira adequada de manipular tais parâmetros.

Ligada às questões de realismo, interatividade e adaptabilidade, uma pesquisa futura interessante 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 fala do sistema pode se comportar de uma forma diferente, através de mudanças em sua entonação. Por outro lado, se o sistema estiver sendo usado em uma vídeo-conferência, mecanismos de clonagem facial [LeGM00] podem buscar fazer com que a face se aproxime das características do usuário interlocutor. Como alternativa para a inclusão dessas funcionalidades, pretende-se investigar o uso do padrão MPEG-4 como uma nova possibilidade para especificação da animação facial no "Expressive Talking Heads".

Outros pontos interessantes de serem abordados em trabalhos futuros são: acrescentar ao sistema novos idiomas para a síntese da voz, principalmente o português; e a alternativa de operar com vozes capturadas. Esse último ponto exigirá, no entanto, a integração com ferramentas de reconhecimento de voz, que, se por um lado podem ampliar o número de aplicações usuárias, por outro podem comprometer a interatividade em tempo real que o sistema oferece.

Uma outra vertente de trabalhos futuros consiste em dar continuidade ao desenvolvimento de aplicações que utilizam o ETHs. Recentemente, foi iniciada a integração do sistema a um formatador hipermídia, o formatador HyperProp [SoMR00]. A idéia é tornar o "Expressive Talking Heads" uma ferramenta de exibição do sistema, participando na exibição de conteúdo em apresentações multimídia/hipermídia. Com a integração, animações faciais para o ETHs podem estar sincronizadas com outros objetos de mídia, compondo, por exemplo, aulas virtuais. Nesse caso, a face do "Expressive Talking Heads" pode inclusive funcionar como uma alternativa a vídeos de professores para clientes que tenham limitações de banda passante na recepção do conteúdo do documento. Um passo seguinte interessante seria introduzir neste sistema integrado a capacidade de geração semi-automática de enredos [CiFF02]. Com isso, seria possível estabelecer uma infra-estrutura para a criação de várias classes de aplicações ricas e interativas, em especial aplicações Web e de TV Digital. Por fim, as questões de emoções ligadas a modelos de inteligência artificial deverão ser tratadas em trabalhos futuros.

Maiores detalhes do trabalho apresentado neste artigo e um vídeo de demonstração do sistema podem ser encontrados em http://www.telemidia.puc-rio.br/~pslr.

Referências Bibliográficas

[BrCS97] Bregler, C.; Covell, M. e Slaney, M. (1997) "Video Rewrite: Driving visual speech with audio". *SIGGRAPH'97*, Los Angeles, EUA.

[CiFF02] Ciarlini, A.; Feijó, B. e Furtado, A. (2002) "An Integrated Tool for Modelling, Generating e Exhibiting Narratives". *2002 AI, Simulation and Planing In High Autonomy Systems*, Lisboa, Portugal, pp. 150-154.

[DiP01] DiPaola, S. (2001) "Facade – The Stanford Facial Animation System". *Technical Report*, Stanford University, EUA, http://dipaola.org/stanford/facade/, visitado em Janeiro de 2002.

[Dut96] Dutoit, T. et al. (1996) "The MBROLA Project: Towards a set of high quality speech synthesis free of use for non-commercial purposes". ICSLP'96, Bélgica.

[Dut97a] Dutoit, T. et al. (1997) "The MBROLA Project", http://tcts.fpms.ac.be/synthesis/mbrola.html, visitado em Maio de 2001.

[Dut97b] Dutoit, T. et al. (1997) "A Short Introduction to Text-to-Speech Synthesis", *Technical Report*, TTS Research Team, TCTS Lab, Faculté Polytechnique de Mons, Bélgica, http://tcts.fpms.ac.be/synthesis/introtts.html, visitado em Maio de 2001.

[EzPo98] Ezzat, T. e Poggio, T. (1998) "MikeTalk: A Talking Facial Display Based on Morphing Visemes", *IEEE Computer Animation*, Center for Biological & Computational Learning and the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Philadelphia, CA, EUA.

[EzPo99] Ezzat, T. e Poggio, T. (1999) "Visual Speech Synthesis by Morphing Visemes", *Technical Report 1658*, Center for Biological & Computational Learning and the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, EUA, C.B.C.L. Paper No. 173.

[GaMa01] Gachery, S. e Magnenat-Thalmann, N. (2001) "Designing MPEG-4 Facial Animation Tables for Web Applications", *Multimedia Modeling Conference*, Amsterdam, Holanda, pp. 39-56.

[Gamm95] Gamma E. et al. "Design Patterns: Elements of Reusable Object-Oriented Software". *Addison Wesley*, 1995.

[GrCr00] Grinberg, M. e Crivicich Grinberg, A. (2000) "MAGPIE", *Third Wish Software & Animation*, Portland, Oregon, EUA, http://www.thirdwishsoftware.com/magpie.html, visitado em Janeiro de 2002.

[Hos92] Hosom, J-P. (1992) "The CSLU Toolkit: A Platform for Research and Development of Spoken-Language Systems", Center for Spoken Language Understanding (CSLU), OGI Campus, Oregon Health & Science University (OGI/OHSU), visitado em Janeiro de 2002, http://cslu.cse.ogi.edu/toolkit/index.html .

[ISO02] ISO/IEC JTC1/SC29/WG11 N4668 (2002) "Overview of the MPEG-4 Standard". http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm, visitado em Julho de 2003

[LeGM00] Lee, W.; Gu, J. e Magnenat-Thalmann, N. (2000) "Generating Animatable 3D Virtual Humans from Photographs" *Eurographics' 2000*, Computer Graphics Forum, 19(3).

[Lu-Ro02] Lucena Rodrigues, P.S. (2002) "Expressive Talking Heads: Um Estudo de Fala e Expressão Facial em Personagens Virtuais", *Dissertação de Mestrado*, Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Brasil.

[Oste98] Ostermann, J. (1998) "Animation of Synthetic Face in MPEG-4", *Computer Animation*, Philadelphia, EUA, pp. 49-56.

[PaWa97] Parke, F.I. e Waters, F. (1996) "Computer Facial Animation", A K Peters, Ltd., Wellesley, MA, ISBN 1-56881-014-8.

[Per97] Perlin, K. (1997) "Responsive Face", Media Research Lab, New York University, EUA, *Trabalho em colaboração*, http://mrl.nyu.edu/~perlin/demox/Face.html, visitado em Abril de 2001.

[SoRM00] Soares L.F.G., Rodrigues R.F. e Muchaluat-Saade D.C. (2000) "Modeling, Authoring and Formatting Hypermedia Documents in the HyperProp System". *ACM Multimedia Systems Journal*, Springer-Verlag, 8 (2), pp. 118-134.

[W3C03] World-Wide Web Consortium (2003) "Web Services Architecture",W3C Working Draft. Em http://www.w3.org/TR/ws-arch/, visitado em julho de 2003.

[WaTC99a] Watt, A., Taylor, P. e Caley, R. (1999) "The Festival Speech Synthesis System: System Documentation", University of Edinburgh, Bélgica, visitado em Junho de 2001, http://www.cstr.ed.ac.uk/projects/festival/manual/ .

[WaTC99b] Watt, A., Taylor, P. e Caley, R. (1999) "The Architecture of the Festival Speech Synthesis System", University of Edinburgh, Bélgica, visitado em Agosto de 2001, http://www-2.cs.cmu.edu/~awb/papers/ESCA98_arch/ .