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

Paula Salgado Lucena[1]*, **Marcelo Gattass**[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,gattass@inf.puc-rio.br, lvelho@impa.br

Abstract. *This article presents a system called Expressive Talking Heads, which focuses on facial animation with synchronization between speech and facial expressions. With this research we were able to propose a taxonomy for talking head systems.*

Resumo. *Este artigo apresenta um novo sistema denominado "Expressive Talking Heads" que trata da animação facial tendo a fala e as expressões faciais sincronizadas. A partir desta pesquisa foi possível propor uma taxonomia para os sistemas talking head.*

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. 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 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 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*.

A principal contribuição deste trabalho é o desenvolvimento de um sistema *talking head* de tempo real que combina a fala com expressividade facial. A entrada é um texto contendo a fala, anotações de expressividade, idioma e gênero. A saída resultante é

*Dissertação defendida no Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro no ano de 2002 com o apoio do CNPq.

uma animação facial em tempo real do personagem virtual enunciando o texto de entrada com o áudio e os movimentos faciais sincronizados. O sistema foi batizado como "Expressive Talking Heads" e para contemplar o seu desenvolvimento se fez necessária uma pesquisa sobre os principais elementos que compõem um sistema *talking head*: fala, face e animação. Como consequência deste estudo foi possível propor uma taxonomia para os sistemas de animação facial. Foi feita também uma análise dos trabalhos de pesquisa nessa área, bem como uma comparação do "Expressive Talking Heads" com os sistemas relacionados. Mais detalhes podem ser encontrados em [Lucena, 2002].

2. Sistemas *Talking Head*: Elementos Principais e Taxonomia

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.

2.1. Fala

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 [Lucena, 2002]. Na primeira abordagem, voz sintetizada, o áudio é gerado através de um sistema *text-to-speech*. 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.

2.2. 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) [Lucena, 2002]. Em ambas, é possível possuir uma face bidimensional ou tridimensional. Uma vez modelada a face, o passo subsequente 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. Por fim, a expressividade é um elemento capaz de enriquecer um sistema *talking head* [Parke and Waters, 1996] e foi trabalhado de forma especial no "Expressive Talking Heads", objetivando ter através da malha poligonal simples utilizada o máximo de expressividade possível.

2.3. Execução

Assim como as classificações apresentadas para a fala e a face, existem duas abordagens no que diz respeito à forma de execução de um sistema *talking head*: tempo real e *in batch* [Lucena, 2002]. A primeira abordagem, execução em tempo real, caracteriza-se

por ser interativa, tendo a inserção dos dados ocorrendo em paralelo à animação produzida como saída do sistema. Já a segunda abordagem, execução *in batch*, 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.

3. O Expressive Talking Heads

Para o desenvolvimento do "Expressive Talking Heads" foi necessário, a partir dos requisitos do sistema, pesquisar qual seria a melhor abordagem para cada um dos parâmetros definidos na taxonomia proposta. No caso da fala verificou-se que a fala sintetizada seria a abordagem ideal, já que desejavasse ter um sistema onde o usuário interagisse continuamente através de uma entrada textual. Para o parâmetro face verificou-se que a melhor abordagem seria de face sintetizada, em decorrência do uso da malha poligonal permitir 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. Por fim, para o parâmetro forma de execução definiu-se a execução em tempo real como um requisito para o sistema. A Figura 1 ilustra uma visão geral do "Expressive Talking Heads". O sistema foi modularizado objetivando ter um melhor entendimento dos elementos principais e dos subsistemas acoplados no seu desenvolvimento. O restante da seção tece alguns comentários sobre cada um dos componentes principais.

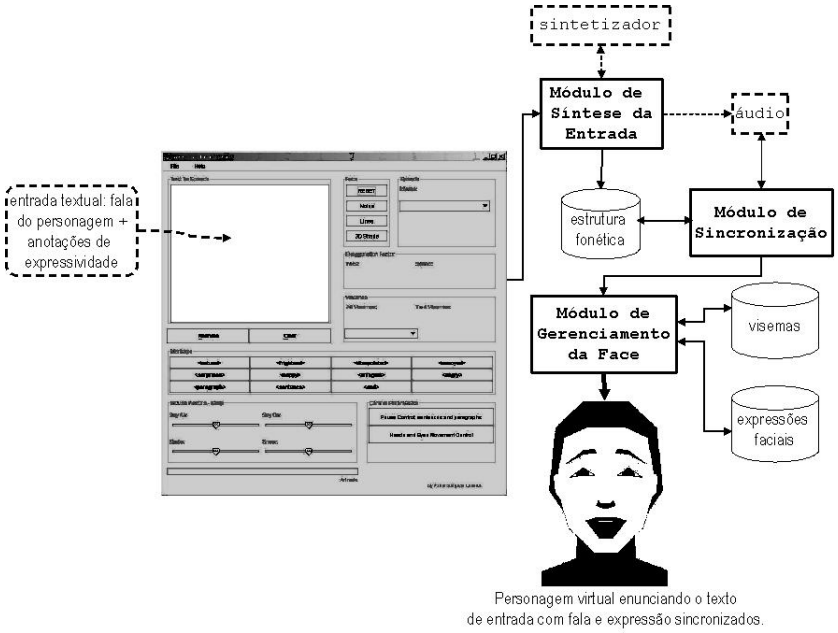

Figura 1: Visão Geral do "Expressive Talking Heads".

3.1. 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 britânico). 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 pelos subsistemas Festival [1] e MBROLA [2] 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 ETHs* acrescenta um tratamento especial à pausa entre as sentenças da fala [Lucena, 2002], a fim de melhorar o realismo do áudio gerado. A comunicação interna entre o Festival e o MBROLA e a integração com o "Expressive Talking Heads" foram feitas através de funções desenvolvidas na linguagem *Scheme*.

3.2. 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 *ResponsiveFace* [3] do qual foi herdada a modelagem da face. Na inicialização do sistema a base de visemas e a base de expressões faciais são carregadas e armazenadas de forma estática. Foram definidos 16 visemas e 8 expressões faciais. Os 16 visemas agrupam os possíveis fonemas existentes, enquanto que as 8 expressões definidas são: natural, assustada, desapontada, incomodada, surpresa, feliz, arrogante e com raiva.

A animação da face durante a fala é feita através da movimentação dos músculos faciais. Estes são construídos através do agrupamento dos vértices que definem a face. Existem 12 vértices, dos quais 5 definem e controlam o movimento dos olhos, 3 músculos definem e controlam as posições da cabeça e 4 músculos definem e controlam os movimentos da boca. Valores dentro do intervalo [-1.0, +1.0] são aplicados a cada músculo determinando quanto o mesmo deve contrair ou relaxar durante a animação.

3.3. 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 propõe animar a face com interatividade e em tempo real. No "Expressive Talking Heads" o controle de sincronização é obtido através do uso de *threads* [Lucena, 2002].

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

[1] Informações e *download* do Festival podem ser encontrados em `http://www.cstr.ed.ac.uk/projects/festival/`

[2] Informações e *download* do MBROLA podem ser encontrados em `http://tcts.fpms.ac.be/synthesis/mbrola.html`

[3] O *ResponsiveFace* é um sistema desenvolvido pelo pesquisador Ken Perlin da NYU. Informações adicionais podem ser encontradas em `http://mrl.nyu.edu/~perlin/demox/Face.html`

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.

4. Conclusões

Animação facial é uma área importante na Computação Gráfica e que está continuamente em crescimento. O trabalho aqui apresentado, em particular, possui três principais contribuições. A primeira delas é a pesquisa dos componentes fundamentais que compõem um sistema *talking head*: fala, face e animação. A segunda é a taxonomia definida para estes sistemas sobre os parâmetros fala, face e forma de execução. Por fim, a terceira e principal contribuição deste trabalho é a implementação do sistema "Expressive Talking Heads", que por sua vez proporcionou contribuições internas. Essas são aqui enumeradas: integração dos subsistemas Festival, MBROLA e *ResponsiveFace*; configuração dos sintetizadores Festival e MBROLA para o trabalho em conjunto, o primeiro funcionando como unidade NLP (*Natural Language Processing*) e o segundo como unidade DSP (*Digital Signal Processing*); combinação de sincronização da fala com expressões; tratamento de pausa objetivando ter um maior realismo na fala; definição de uma linguagem de marcação para a emoção que o personagem deve assumir na fala; definição do gênero da voz e idioma da fala como parâmetros para o usuário; definição de diferentes abordagens para o movimento dos componentes faciais; e, por fim, a definição da base de visemas.

Como trabalhos futuros é possível destacar duas principais vertentes. A primeira é o desenvolvimento de aplicações reais baseadas no "Expressive Talking Heads", tais como *3D chats*, sistemas educacionais e personagens para TV interativa. A segunda vertente é a pesquisa aprofundada e consequente implementação de aspectos de expressividade, tais como incorporação de emoções na fala e o uso de outras abordagens para a fala como a entrada no sistema através de reconhecimento de voz.

Por fim, o endereço *http://www.inf.puc-rio.br/ pslucena/mestrado/* apresenta maiores detalhes a respeito do sistema "Expressive Talking Heads" e um vídeo que ilustra a execução do sistema.

Referências

Lucena, P. S. (2002). Expressive talking heads: Um estudo de fala e expressão facial em personagens virtuais. Dissertação de Mestrado, Departamento de Informática/PUC-Rio.

Parke, F. I. and Waters, K. (1996). *Computer Facial Animation*. A K Peters, Ltd., Wellesley, MA. ISBN 1-56881-014-8.