Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e Computação

Animação facial 2D sincronizada com a fala baseada em imagens de visemas dependentes do contexto fonético

Autora: Paula Dornhofer Paro Costa
Orientador: Prof. Dr. José Mario De Martino

Tese de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Engenharia de Computação.

Banca Examinadora:
Prof. Dr. José Mario De Martino - DCA/FEEC/UNICAMP
Dr. Leandro de Campos Teixeira Gomes - Fundação CPqD
Prof. Dr. Roberto de Alencar Lotufo - DCA/FEEC/UNICAMP

Campinas, SP
Junho/2009

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

C823a

Costa, Paula Dornhofer Paro
Animação facial 2D sincronizada com a fala baseada em imagens de visemas dependentes do contexto fonético/Paula Dornhofer Paro Costa. –Campinas, SP: [s.n.], 2009.

Orientador: José Mario De Martino.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Computação gráfica. 2. Animação.
3. Metamorfose. 4. Animação por computador.
5. Comunicação visual. I. De Martino, José Mario.
II. Universidade Estadual de Campinas.
Faculdade de Engenharia Elétrica e de Computação.
III. Título

Título em Inglês:	Speech synchronized 2D facial animation based on phonetic context dependent visemes images
Palavras-chave em Inglês:	Computer graphics, Animation, Morphing, Computer animation, Visual communication
Área de concentração:	Engenharia de Computação
Titulação:	Mestre em Engenharia Elétrica
Banca Examinadora:	Leandro de Campos Teixeira Gomes, Roberto de Alencar Lotufo
Data da defesa:	22/06/2009

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidata: Paula Dornhofer Paro Costa

Data da Defesa: 22 de junho de 2009

Título da Tese: "Animação Facial 2D Sincronizada com a Fala Baseada em Imagens de Visemas Dependentes do Contexto Fonético"

Prof. Dr. José Mario De Martino (Presidente): _____

Dr. Leandro de Campos Teixeira Gomes: _____

Prof. Dr. Roberto de Alencar Lotufo: _____

Resumo

A animação facial por computador sincronizada com a fala permite a implementação de cabeças virtuais que podem contribuir para tornar interfaces humano-computador mais eficientes e atraentes. O presente trabalho apresenta um método de síntese de animação facial 2D baseado em imagens cujo desenvolvimento foi guiado por dois objetivos principais: a reprodução realista da movimentação articulatória visível da fala, incluindo os efeitos da coarticulação, e a possibilidade de implementação do método mesmo em plataformas com capacidades limitadas de processamento e memória, tais como celulares e assistentes pessoais digitais. O método desenvolvido baseia-se em uma base de imagens de visemas dependentes de contexto para o Português do Brasil e adota a técnica de metamorfose entre visemas para a síntese da animação facial. A abordagem proposta representa uma estratégia de síntese alternativa e inovadora, capaz de reproduzir a movimentação articulatória visível da fala, incluindo os efeitos da coarticulação, a partir de uma base de apenas 34 imagens. O trabalho inclui a implementação de um sistema piloto integrado a conversor texto-fala. Adicionalmente, o método de síntese proposto é avaliado através de teste de inteligibilidade da fala. Os resultados desta avaliação indicam que a informação visual fornecida pelas animações geradas pelo sistema contribuem para a inteligibilidade da fala em condições de áudio contaminado por ruído. Apesar do trabalho estar restrito ao Português do Brasil, a solução apresentada é aplicável a outras línguas.

Palavras-chave: Computação Gráfica, Animação Facial, Visemas, Coarticulação, Metamorfose.

Abstract

Speech synchronized facial animation allows the implementation of talking heads that potentially can improve human-computer interfaces making them more efficient and attractive. This work presents an image based 2D facial animation synthesis method whose development was guided by two main goals: the realistic reproduction of visible speech articulatory movements, including coarticulation effects, and the possibility to implement the method also on limited processing and memory platforms, like mobile phones or personal digital assistants. The developed method is based on an image database of Brazilian Portuguese context dependent visemes and uses the morphing between visemes strategy as facial animation synthesis technique. The proposed approach represents an alternative and innovative synthesis strategy, capable of reproducing the visible speech articulatory movements, including coarticulation effects, from an image database of just 34 images. This work includes the implementation of a pilot system integrated to a text-to-speech synthesizer. Additionally, the proposed synthesis method is evaluated through a speech intelligibility test. The test results indicate that the animations generated by the system contribute to improve speech intelligibility when audio is degraded by noise. Despite the fact this work is restricted to Brazilian Portuguese, the presented solution is applicable to other languages.

Keywords: Computer Graphics, Facial Animation, Visemes, Coarticulation, Morphing.

Agradecimentos

Ao meu orientador, Prof. Dr. José Mario De Martino, pela confiança depositada e pela liberdade concedida durante o desenvolvimento deste trabalho, que serviram de estímulo à criatividade e me mantiveram sempre motivada. Pela sua tolerância, paciência e objetividade nos momentos de redirecionamento, que resultaram em grandes oportunidades de crescimento pessoal e profissional.

Aos Profs. Drs. Clesio Luis Tozzi, Leo Pini Magalhães e Roberto de Alencar Lotufo, pelas valiosas sugestões.

A Norberto Alves Ferreira da *Gerência de Serviços e Aplicações Multimídia* da Fundação CPqD, por apoiar o desenvolvimento deste trabalho através da disponibilização de recursos financeiros, humanos e de infra-estrutura.

À Fundação CPqD pelo apoio financeiro e iniciativa de incentivo à pesquisa.

À equipe responsável pelo desenvolvimento e suporte do "CPqD Texto Fala", pelo entusiasmo em relação ao projeto, pelas idéias inspiradoras, pelas inúmeras dúvidas solucionadas e, acima de tudo, pela amizade. Em particular, agradeço a Edson José Nagle pelas horas de atenção dedicadas ao trabalho, em discussões enriquecedoras, através das quais pude compartilhar de sua vasta experiência.

A Carolina Franciscangelis, pela sua disponibilidade em colaborar com este trabalho e pela seriedade com que desempenhou o papel de apresentadora durante a captura do corpus audiovisual.

À equipe técnica do laboratório da *Gerência de Serviços e Aplicações Multimídia* da Fundação CPqD, pelo suporte prestado durante a captura do corpus audiovisual.

A Anderson Luiz Brunozi, Isidro Lopes da Silva Neto e Vinicius de Lima pelo desenvolvimento da aplicação de demonstração do sistema piloto de animação facial.

Aos colaboradores da Fundação CPqD que gentilmente participaram da avaliação do sistema piloto de animação facial.

A José Augusto Ribeiro e Franklin César Flores, pela colaboração no desenvolvimento de algoritmos de processamento digital de imagens.

À minha família, pelo apoio e motivação essenciais ao desenvolvimento deste trabalho.

Ao meu querido pai que, dentre tantas outras coisas, me ensinou o valor do estudo e da autodisciplina.

À minha grande amiga Melanie, mãe de verdade, mulher inteligente e admirável, que nos ensinou a gostar de matemática e sempre nos dá ideias de invenções geniais.

Ao meu irmão, que já se tornou um grande homem, mas que sempre será Pedrinho em meu coração.

Ao meu querido marido Rodrigo, que compartilha de minhas paixões, meu companheiro de vida e de profissão.

Às minhas filhas Ana Luiza e Laura, que trouxeram este amor imenso à minha vida e que através de seus brilhantes e sagazes olhares infantis me fazem vislumbrar um mundo cada vez melhor.

Sumário

Lista de Figuras

Lista de Tabelas

Trabalhos Publicados Pelo Autor

1. DE MARTINO J. M.; COSTA, P. D. P. “Sistema de síntese de animação facial por computador baseada na manipulação de imagens”. Depósito do pedido junto ao INPI: 13 de maio de 2009.

2. COSTA, P. D. P.; DE MARTINO, J. M.; NAGLE, E. J. “Speech Synchronized Image-Based Facial Animation”. In: *Proceedings of the International Workshop on Telecommunications 2009 - IWT 2009*. São Paulo, Brazil, p. 235-241, February 2009.

3. COSTA, P. D. P.; DE MARTINO, J. M.; NAGLE, E. J. “Sistema de animação facial 2D sincronizado com a fala integrado ao CPqD Texto Fala”. In: *Anais do I Congresso Tecnológico Infobrasil - Infobrasil 2008*. Fortaleza, CE, Brasil: [s.n.], 2008.

Capítulo 1

Introdução

1.1 Introdução

A recente expansão mundial das tecnologias de comunicação móvel, acompanhada pela maior disponibilização e aumento de capacidade das redes de dados, tornou possível o acesso ubíquo à informação e trouxe à tona novas necessidades e possibilidades de aplicações.

Por outro lado, a consolidação de padrões de interfaces e protocolos de comunicação e a evolução tecnólogica de microprocessadores, microcontroladores e dispositivos de memória, possibilitaram o surgimento de um grande número de dispositivos portáteis que, entre outras funcionalidades, proveem acesso a redes de dados. Como consequência, observou-se uma expansão extraordinária no número de usuários que fazem uso destes novos e diferentes meios de acesso à informação e que não necessariamente possuem familiaridade com interfaces tradicionalmente utilizadas em computadores pessoais do tipo WIMP (*Windows, Icons and Pointing Devices*), onde os mecanismos de entrada/saída são normalmente caracterizados por teclado, mouse e monitor.

Neste contexto, a pesquisa voltada para o desenvolvimento de agentes humanos virtuais interativos surge como uma alternativa emergente e promissora para a implementação de interfaces e aplicações mais intuitivas e humanizadas. A partir de esforços multidisciplinares de pesquisa e desenvolvimento nas áreas de reconhecimento de voz, processamento da linguagem natural, inteligência artificial, síntese da fala, computação gráfica e animação, é possível implementar personagens virtuais capazes de capturar mais facilmente a atenção do usuário e tornar a atividade de interação mais atrativa e envolvente (GRATCH et al., 2002). Dependendo da aplicação, tais agentes podem desempenhar papéis variados, tais como instrutores, assistentes, avatares, apresentadores, atendentes ou vendedores.

Com esta visão, este trabalho dedica-se ao desenvolvimento da tecnologia de animação facial sincronizada à fala como componente essencial para a implementação de cabeças virtuais personificadas, ou *talking heads*.

O desenvolvimento de *talking heads* leva em consideração o papel de destaque que a face ocupa na comunicação humana. Desde o nascimento somos treinados nos mecanismos de comunicação face a face e, estimulados por experiências sociais, nos tornamos capazes de interpretar e identificar estados emocionais transmitidos pela face, utilizando sua informação visual para complementar a compreensão da mensagem transportada pelo sinal acústico da fala.

Neste sentido, um dos objetivos da animação facial gerada por computador é conferir a uma face

virtual a aparência, a movimentação e o comportamento de uma face real. Esta capacidade pode ser qualitativamente expressa em termos do grau de vídeo-realismo alcançado pela animação, ou seja, sua capacidade de ser confundida com o vídeo de uma face real. Assim, uma animação facial vídeo-realista, além da reprodução fotográfica das características estáticas da face (como rugas e textura da pele), é também capaz de reproduzir os movimentos articulatórios da fala em sincronia e harmonia com a locução. Além disso, também são importantes a reprodução de gestos de comunicação não verbais (como expressões faciais ou o menear de cabeça) e movimentos fisiológicos não relacionados com a comunicação (como o piscar de olhos para lubrificação dos mesmos).

A reprodução realista dos movimentos articulatórios da fala é obtida levando-se em consideração os mecanismos de produção da mesma. A realização acústica dos diversos fonemas de uma língua se dá através de configurações típicas do trato vocal que, entre outros elementos articuladores, inclui as cordas vocais, o palato, a cavidade nasal, a língua e os lábios. No entanto, apenas uma parcela dos movimentos realizados pelos órgãos articuladores é visualizada na face através, principalmente, da movimentação dos lábios e da região em torno deles. Assim, a modelagem dos movimentos articulatórios faciais visíveis pode ser realizada através de visemas. Neste trabalho, um visema é definido como uma postura labial estática que é visualmente contrastiva a outra e que pode ser associada à realização acústica de um fonema. É importante ressaltar que a realização acústica dos fonemas não é realizada de maneira isolada, mas envolve efeitos de coarticulação. Os efeitos da coarticulação se manifestam pela alteração do padrão articulatório de um determinado segmento sonoro pela influência de outro adjacente, ou próximo, na cadeia de produção sonora. Os efeitos da coarticulação fazem com que, por exemplo, o "p" da palavra "paro" seja visualmente distinto do "p" da palavra "puro". Neste último caso, o movimento articulatório necessário à produção do "u" influencia de maneira significativa os aspectos visíveis da articulação do "p".

Neste trabalho, os efeitos da coarticulação são contemplados a partir de uma abordagem direta e eficiente por meio da utilização de visemas dependentes de contexto identificados para o Português do Brasil por DE MARTINO (2005). No contexto deste trabalho, visemas dependentes de contexto são definidos como posturas labiais estáticas associadas não somente à produção de um segmento isolado, mas influenciadas por segmentos específicos que o antecedem e sucedem em uma sequência de locução.

Outro aspecto importante para a obtenção de uma animação facial vídeo-realista é a abordagem adotada para modelagem da face humana. Este trabalho apresenta um processo de síntese de animação facial 2D em que a modelagem da face é baseada em imagens. O processo apresentado manipula imagens fotográficas extraídas de um corpus audiovisual obtido a partir do vídeo de uma face real, que constituem uma base de imagens. A síntese da animação é implementada tendo-se como parâmetro de entrada a transcrição fonética temporizada da fala a ser visualmente animada. A partir das informações fornecidas pela transcrição fonética, a animação é sintetizada através do apropriado sequenciamento, concatenação e apresentação de quadros resultantes do processamento de imagens da base.

O processo de síntese da animação apresentado neste trabalho implementa a abordagem de animação facial 2D a partir de uma base de imagens de visemas dependentes de contexto. A dinâmica da animação é obtida a partir da técnica de metamorfose entre imagens. Tal abordagem permite a implementação de um sistema de animação facial 2D capaz de contemplar os efeitos da coarticulação a partir de uma base de apenas 34 imagens. Com isso, o sistema pode ser adaptado a plataformas de capacidade de processamento e memória limitados como telefones celulares, PDAs (*Personal Digital*

Assistants) e decodificadores de TV digital.

O trabalho também abrange a implementação de um sistema piloto em que a fala é gerada por um sistema de conversão texto-fala para o Português do Brasil, que gera o aúdio e fornece ao sistema de animação a transcrição fonética temporizada correspondente à locução anteriormente especificada através de uma entrada textual. A implementação piloto permitiu a avaliação das animações geradas pelo processo de síntese apresentado, cujos resultados são também apresentados neste trabalho.

A principais contribuições deste trabalho podem ser resumidas em:

- a implementação de um sistema de animação facial 2D para o Português do Brasil;
- o desenvolvimento de um processo de síntese de animação facial 2D baseado na metamorfose entre visemas dependentes de contexto capaz de contemplar os efeitos da coarticulação a partir de uma base de imagens reduzida (34 imagens);
- o estabelecimento de uma abordagem de síntese da animação que pode ser adaptada a dispositivos com capacidade limitada de processamento e armazenamento de dados;
- a obtenção de um sistema de síntese de animação facial capaz de sintetizar animações com nível de vídeo-realismo e favorecer a inteligibilidade da fala em condições desfavoráveis de áudio.

Vale destacar que, apesar do sistema ter sido implementado para o português do Brasil, os princípios básicos que regem o sistema podem ser aplicados a qualquer língua.

O presente trabalho é organizado da seguinte maneira:

- **Capítulo 1 - Introdução**

  Neste capítulo são apresentados: a motivação para o desenvolvimento deste trabalho, as principais características da solução apresentada, as principais contribuições e a organização do trabalho.

- **Capítulo 2 - Animação Facial 2D**

  A proposta deste capítulo é realizar uma breve comparação entre a animação facial 3D e 2D, ressaltando os motivos da escolha da abordagem 2D. Em seguida, realiza-se a revisão das abordagens existentes de animação facial 2D com base na literatura associada. A partir desta revisão é derivada a taxonomia dos sistemas existentes, explicitando-se a abordagem adotada por este trabalho.

- **Capítulo 3 - Construção da Base de Imagens**

  Neste capítulo são descritas as diretrizes e o processo de captura do corpus audiovisual e as etapas de pré-processamento das imagens para construção da base de imagens.

- **Capítulo 4 - Síntese da Animação Facial**

  Este capítulo descreve como as informações da transcrição fonética temporizada são utilizadas para determinar a sequência de quadros da animação. Em seguida, são fornecidos os detalhes da síntese da animação através da metamorfose entre visemas.

- **Capítulo 5 - Implementação Piloto do Sistema de Animação Facial 2D**

  Descrição do sistema piloto implementado, incluindo a integração realizada com o sistema de conversão texto-fala "CPqD Texto Fala". O capítulo apresenta ainda os principais aspectos relativos à implementação do software de síntese e realiza uma análise do tempo de síntese de animação apresentado pelo sistema.

- **Capítulo 6 - Teste de Inteligibilidade da Fala**

  Descrição da avaliação aplicada ao sistema através da realização de testes de inteligibilidade da fala. O capítulo apresenta a metodologia empregada para a realização dos testes, tratamento estatístico realizado com os dados, os resultados e as conclusões sobre os mesmos.

- **Capítulo 7 - Conclusões**

  Capítulo final onde são repassadas as principais contribuições e os desenvolvimentos futuros estimulados pelo trabalho.

Capítulo 2

Animação Facial 2D

2.1 Introdução

A tecnologia de animação facial por computador tem sido objeto de pesquisa desde a década de 70, tendo como importante marco inicial o trabalho pioneiro de Parke (1972). Desde então, as técnicas de animação facial por computador experimentaram uma rápida evolução expressa pelas diferentes abordagens propostas na literatura e pela obtenção de animações cada vez mais realistas e convincentes (NOH; NEUMANN, 1998).

Dentre os fatores que impulsionaram tal evolução, podem-se destacar:

- a crescente busca por interfaces humano-computador mais naturais, eficientes e atrativas devido ao surgimento de novas aplicações e modalidades interativas de acesso à informação;
- a popularização de sistemas computacionais com alta capacidade de processamento e armazenamento de dados;
- o desenvolvimento de técnicas de processamento de imagens e dados que servem de suporte à implementação de sistemas de animação, destacando-se os avanços observados na área de visão computacional e a implementação de algoritmos computacionalmente mais eficientes e viáveis voltados para a análise de grandes volumes de dados multidimensionais.

Paralelamente aos avanços tecnológicos da animação facial, os sistemas de síntese da fala alcançaram no mesmo período a capacidade de sintetizar sinais de fala de qualidade próximos da fala humana (NG, 1998). Tal fato permitiu a concepção e implementação de sistemas de animação facial sincronizados à fala, ou *talking heads*, que rapidamente mostraram seu potencial de aplicação em áreas como entretenimento e educação, bem como na implementação de agentes virtuais no papel de assistentes, apresentadores, agentes sociais, avatares, etc. (COSATTO et al., 2003).

Este capítulo apresenta uma visão geral sobre a implementação de *talking heads* baseadas em sistemas de animação facial 2D através da apresentação do estado da arte para esta técnica. Um dos objetivos deste capítulo é apresentar o contexto e introduzir os conceitos que serviram de base para a implementação deste trabalho. Para isso, são analisadas as principais linhas de trabalho encontradas na literatura para sistemas *talking heads*, tendo-se como foco principal o contexto 2D.

Inicialmente, na Seção 2.2, apresentam-se a definição e as diferenças encontradas em sistemas de animação facial baseados em modelo, ou 3D, e sistemas de animação baseados em imagens, ou 2D. Em seguida, na Seção 2.3, descreve-se o processo de modelagem da face por intermédio de uma base de imagens e a etapa de síntese de *talking heads* baseadas na abordagem 2D. A Seção 2.4 expõe uma classificação destes sistemas com base na estratégia de síntese adotada apresentando-se os trabalhos mais significativos para cada abordagem apresentada. Os comentários finais do capítulo são apresentados na Seção 2.5.

2.2 Sistemas de Animação Facial

Os sistemas de animação facial atualmente existentes podem ser divididos em duas correntes principais de desenvolvimento que se diferenciam pela abordagem adotada na modelagem da cabeça humana e seus elementos: a animação baseada em modelo, ou 3D, e a animação baseada em imagens, ou 2D.

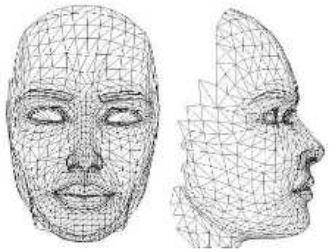
Na abordagem 3D, a cabeça e seus elementos são modelados através de um modelo geométrico, via de regra tridimensional, descrito, por exemplo, por malhas poligonais (Figura 2.1(a)). A este modelo tipicamente são associadas imagens de texturas que visam reproduzir a aparência visual de elementos como cabelos, pele e rugas. Nesta abordagem, a reprodução da movimentação articulatória visível da fala é realizada através de sofisticados modelos de manipulação da geometria e de adequação das texturas utilizadas. Por este motivo, a síntese de animações 3D de aspecto natural, realistas e convincentes está associada a um elevado custo computacional de implementação, geralmente acompanhado de laboriosos ajustes particulares à aplicação de destino. O trabalho de Parke e Waters (1996) fornece uma introdução a esta abordagem, abrangendo os seus diversos aspectos.

A animação facial 2D é obtida através do apropriado sequenciamento, concatenação e apresentação de imagens fotográficas de uma face real (Figura 2.1(b)). Nesta abordagem, a modelagem da face é realizada por intermédio de uma base de imagens construída a partir da análise e processamento de um corpus audiovisual de um apresentador real. Tendo-se como foco a reprodução da movimentação articulatória visível da fala, as imagens armazenadas na base são correspondentes a visemas, imagens de posturas labiais visualmente contrastantes entre si associadas aos diversos sons da língua [1]

O tamanho da base de imagens e as diretrizes utilizadas para sua constituição são os principais aspectos envolvidos na modelagem 2D, enquanto a síntese de animações vídeo-realistas depende da apropriada seleção de imagens da base e algoritmos de processamento de imagens para a composição da animação final (BREGLER; COVELL; SLANEY, 1997), (EZZAT; POGGIO, 1998), (COSATTO; GRAF, 1998), (EZZAT; GEIGER; POGGIO, 2002).

Ao contrário da abordagem 3D, a animação facial 2D não permite facilmente a reprodução de amplos movimentos de rotação e mudança de pose da cabeça uma vez que a informação contida nas imagens é puramente bidimensional. Por outro lado, estes sistemas são denominados inerentemente fotorrealistas devido à natureza fotográfica das imagens utilizadas na síntese da animação. Considerando-se estes aspectos, sistemas de animação facial 2D tipicamente são implementados com menor custo computacional devido à menor complexidade dos modelos utilizados durante a síntese.

[1] O termo *viseme* (traduzido para visema neste trabalho) foi cunhado por Fisher (1968) como resultado da contração da palavra “visual” and “phoneme”. Na literatura, encontram-se variações da definição do termo “visema”. Neste trabalho, visema é a realização visual estática da postura articulatória característica de um segmento sonoro da fala.



(a) Modelo 3D

(b) Sequência de imagens fotográficas

Fig. 2.1: Modelagem da face segundo as abordagens 3D e 2D

Vale destacar a combinação das abordagens 3D e 2D em sistemas em que a fronteira entre as duas vertentes de desenvolvimento não são tão nítidas. Em sistemas como (COSATTO; GRAF, 2000), (BROOKE; SCOTT, 1998), (PIGHIN et al., 1998) e (KSHIRSAGAR; MAGNENAT-THALMANN, 2003), a reprodução dos movimentos da face é obtida pela combinação das estratégias de manipulação de um modelo geométrico tridimensional e a análise e processamento de imagens fotográficas de diferentes expressões faciais e/ou posturas labiais capturadas de uma face real. A combinação dessas técnicas em geral confere à animação facial maior grau de liberdade na movimentação rígida da cabeça e maior nível de vídeo-realismo.

2.3 Sistemas de Animação Facial 2D

A implementação de *talking heads* baseadas na abordagem 2D tem como ponto de partida o projeto e construção de uma base de imagens e informações auxiliares utilizadas pelo processo de síntese da animação facial. As diretrizes empregadas para construção de uma base de imagens, por sua vez, refletem importantes aspectos da estratégia adotada para o processo de síntese da animação facial 2D (BREGLER; COVELL; SLANEY, 1997), (EZZAT; POGGIO, 1998), (COSATTO; GRAF, 2000), (EZZAT; GEIGER; POGGIO, 2002), (EDGE; MADDOCK, 2003).

Esta seção descreve, com base na literatura, as principais abordagens para a implementação de uma base de imagens e para o processo de síntese de animações faciais 2D.

2.3.1 Base de Imagens

Para a contrução da base de imagens, inicialmente são definidas as características de um corpus audiovisual do qual serão extraídas as amostras que formarão esta base de imagens. Os principais

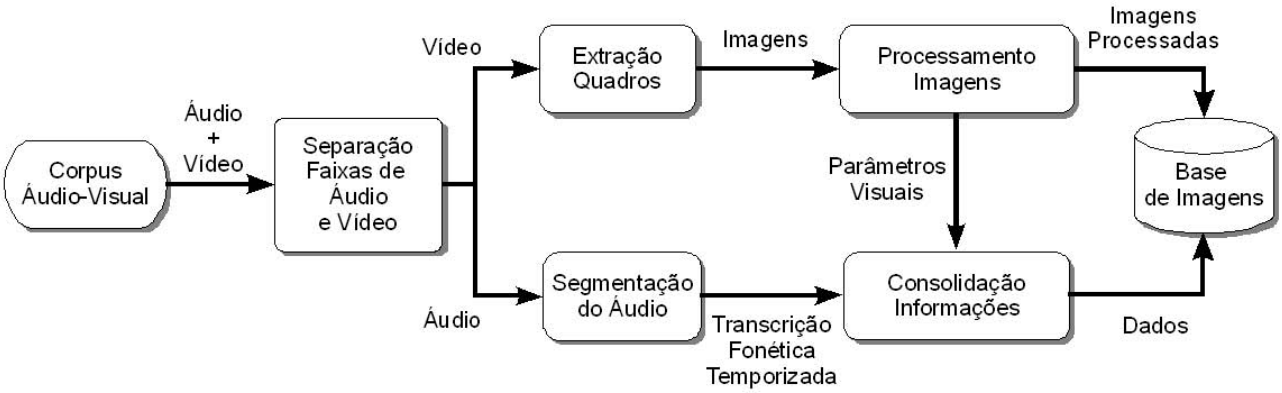


Fig. 2.2: Construção de uma Base de Imagens em Sistemas de Animação Facial 2D

aspectos envolvidos são: o cenário, a apresentação visual do locutor e o conteúdo pronunciado por ele. Assim, o processo de captura do corpus audiovisual pode variar desde a escolha de um trecho de vídeo qualquer sem imposição de restrições (BREGLER; COVELL; SLANEY, 1997) a um complexo processo de captura audiovisual realizado sob condições controladas de cenário (fundo, iluminação, som, equipamentos, etc.) em que um ator apresenta aparência e postura pré-definidas e seu conteúdo de locução é previamente estabelecido.

Uma vez definido o corpus audiovisual, segue-se uma série de operações de pré-processamento que podem ser tipicamente resumidas pela Figura 2.2.

A primeira operação é a separação das faixas de áudio e vídeo do material audiovisual capturado.

O sinal de áudio passa pelo processo de segmentação de fones que resultará na transcrição fonética temporizada. Este processo pode ser totalmente manual, semi-automático (através do fornecimento da transcrição do texto pronunciado) ou totalmente automático através da aplicação de algoritmos de reconhecimento da fala.

A faixa de vídeo passa por um processo de segmentação a partir do qual são extraídas imagens correspondentes a quadros de vídeo. Estas imagens por sua vez são, via de regra, digitalmente processadas visando a uniformização e correção de pose da cabeça nas imagens, regularização de iluminação e extração de parâmetros visuais das imagens.

Finalmente, as informações obtidas pela transcrição fonética temporizada do áudio são confrontadas com as informações extraídas das imagens. Realiza-se assim um processo de consolidação das informações em que cada imagem é devidamente rotulada e associada à produção de um determinado fone em um determinado contexto de locução, num processo de identificação de visemas.

A construção de uma base de imagens em sistemas 2D caracteriza o processo de modelagem visual da face, incluindo informações que podem ou não estar associadas à modelagem de reprodução dos movimentos articulatórios fala. O número de imagens extraídas do corpus audiovisual e as informações a elas associadas estão intimamente ligadas à abordagem de síntese a ser implementada.

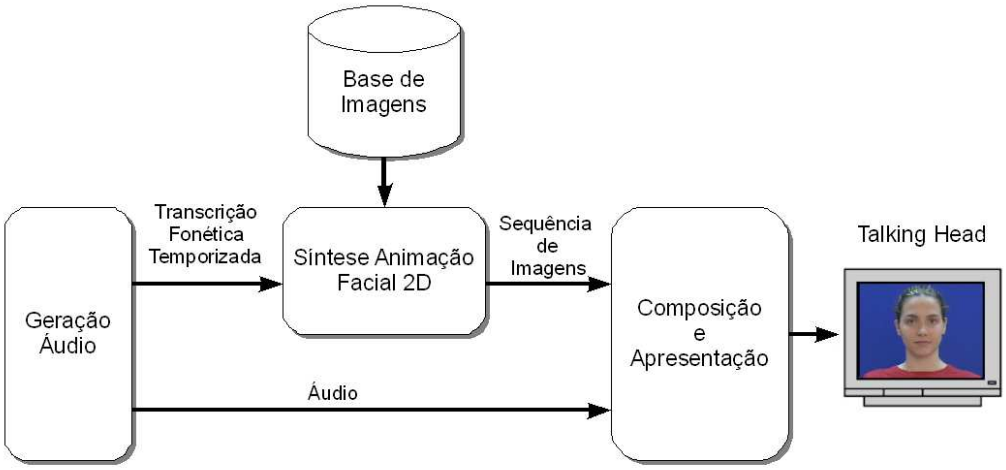


Fig. 2.3: Síntese em Sistemas de Animação Facial 2D

2.3.2 Síntese da Animação Facial 2D

A síntese em sistemas de animação facial corresponde ao processo de geração de uma animação em sincronia e harmonia com o áudio correspondente à fala a ser visualmente animada. A síntese de *talking heads* baseadas na abordagem 2D pode ser tipicamente resumida pela cadeia de processos mostrada na Figura 2.3.

O processo de síntese inicia-se com a definição do áudio correspondente à fala, que pode ser resultado de um processo de gravação da voz de um locutor ou de um sistema de síntese da fala. Apesar da captura de voz tender a proporcionar efeitos mais realistas do que a voz sintetizada, os atuais sistemas de síntese de fala produzem excelentes resultados e são mais flexíveis, uma vez que a síntese de um novo conteúdo é um processo inteiramente automatizado.

Uma abordagem bastante empregada em sistemas *talking heads* é a geração do áudio da fala através de sistemas do tipo conversor texto-fala (TTS - *Text to Speech*) capazes de gerar um sinal de fala a partir de um texto fornecido como entrada.

Independentemente da abordagem utilizada para obtenção do áudio correspondente à fala, a maioria dos sistemas de animação facial 2D possuem como parâmetro de entrada a transcrição fonética temporizada da fala a ser animada. A transcrição fonética temporizada é composta pela sequência de fones que compõem a locução e suas respectivas durações. Ela pode ser obtida através de processos como: análise manual do áudio gerado, aplicação de algoritmos de segmentação automática da voz ou pode ser um resultado intermediário do processo de conversão texto-fala.

As informações fornecidas pela transcrição fonética temporizada são utilizadas pelo sistema para selecionar imagens da base de imagens e, após processá-las, gerar uma sequência de quadros que reproduzem na face virtual a movimentação articulatória visível. O processamento das imagens da base visa garantir transições naturais e suaves entre visemas ou entre sequências de quadros.

Vale citar sistemas que não utilizam a transcrição fonética temporizada como informação intermediária entre o áudio da fala e o sistema de animação facial. Em (BRAND, 1999) e (BASU et al.,) os critérios de seleção de visemas da base são determinados a partir da análise do sinal de áudio que fornece parâmetros de entrada a um modelo estatístico previamente definido.

2.4 Classificação de Sistemas de Animação Facial 2D

Considerando-se a estratégia para o processamento, geração e concatenação de quadros da animação final, as técnicas de síntese de animação facial 2D podem ser divididas em 3 abordagens principais (Figura 2.4):

- Metamorfose entre visemas;
- Síntese concatenativa;
- Trajetória de visemas em espaço multidimensional.

A seguir estas técnicas são analisadas e comparadas levando-se em consideração os seguintes aspectos principais:

- natureza do corpus audiovisual;
- estratégia adotada para construção da base de imagens;
- modelagem adotada para reprodução da movimentação articulatória;
- algoritmos de seleção de imagens da base;
- estratégia de processamento, geração e concatenação de quadros da sequência final de animação.

Fig. 2.4: Classificação das Técnicas de Síntese de Sistemas de Animação Facial 2D

2.4.1 Metamorfose entre visemas

A síntese visual da fala baseada na metamorfose entre visemas pode ser considerada, historicamente, a primeira técnica utilizada para implementação de sistemas de animação facial 2D (SCOTT et al., 1994), tendo suas origens na abordagem convencional de animação baseada na interpolação entre poses-chave (*key-framing*) (PARKE; WATERS, 1996).

Nesta técnica, cada fonema da língua é representado por uma única imagem, representando um visema, da base de imagens. A partir da transcrição temporizada da sequência de fones, o processo

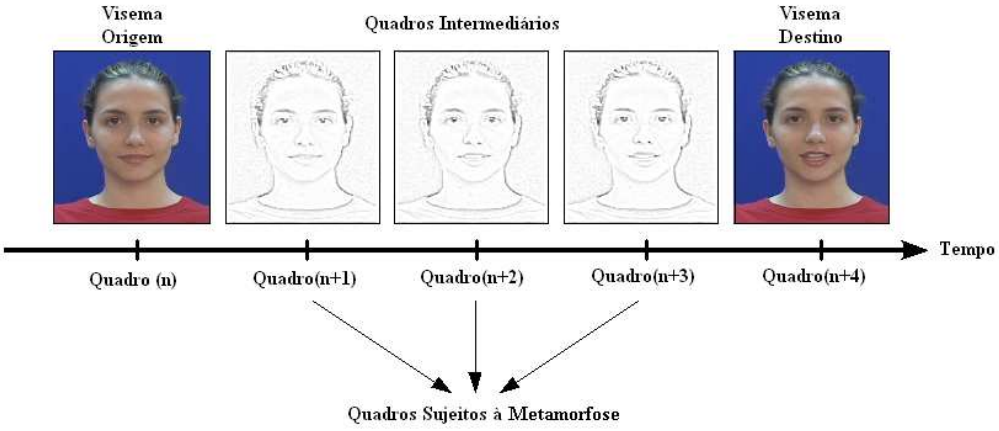


Fig. 2.5: Síntese por Metamorfose entre Visemas

de síntese associa imagens de visemas a diferentes instantes da animação, definindo quadros-chaves. Os quadros intermediários entre os quadros-chave são gerados com a técnica de metamorfose entre imagens.

A metamorfose entre imagens é uma técnica de processamento digital tipicamente utilizada como ferramenta para geração de transições suaves e fluidas entre duas imagens visualmente contrastantes entre si. O processo de metamorfose inicia-se com a definição de um mapa de correspondência entre pontos-âncora previamente definidos na imagem inicial (visema de origem) e na imagem final (visema de destino), que garantem a preservação da correspondência desejada entre atributos geométricos das imagens a serem transformadas. Em sistemas de animação facial, os pontos-âncora em geral são associados a elementos como olhos, nariz, orelhas, lábios e contorno do queixo, que melhor definem as características geométricas de uma face. É possível a generalização do conceito de pontos-âncora abrangendo características compostas por conjuntos de pontos tais como linhas ou regiões nas imagens, como mostrado na Figura 2.6.

Após a definição dos pontos-âncora e suas correspondências nas duas imagens, uma função de transformação geométrica espacial (*warping*) é aplicada ao visema de origem, distorcendo-o gradualmente em direção ao visema destino. Analogamente, o *warping* é aplicado no sentido inverso, distorcendo-se o visema de destino em direção ao visema de origem. A metamorfose é concluída combinando-se os resultados das duas distorções espaciais em sentidos opostos através de uma interpolação em função do tempo dos valores de cores dos pixels das imagens obtidas. Wolberg (1998) apresenta conceitos e estratégias associadas à técnica de metamorfose de imagens.

Um primeiro trabalho que aplica a estratégia de metamorfose entre visemas é descrito em (SCOTT et al., 1994). Nesse trabalho, os autores descrevem a implementação de um sistema de animação facial com uma base de imagens contendo 50 visemas associados a diferentes fonemas da língua inglesa e capturados a partir de um corpus audiovisual filmado em condições controladas. A principal contribuição desse trabalho é o pioneirismo da solução apresentada. No entanto, o sistema não contempla a modelagem dos efeitos da coarticulação, além de não ser detalhado o algoritmo de metamorfose empregado.

Em (EZZAT; POGGIO, 1998) é apresentado o sistema *MikeTalk*. Este sistema tem como principal característica a utilização de uma base de imagens extremamente reduzida, contendo apenas 16

(a) Mapa de correspondência de pontos (extraído de (EDGE; MADDOCK, 2003))

(b) Mapa de correspondência de linhas (adaptado de (WOLBERG, 1998))

Fig. 2.6: Exemplos de mapas de correspondência entre imagens

visemas extraídos de um corpus audiovisual capturado em condições controladas. O número reduzido de imagens é resultado da associação de um único visema a grupos de fonemas que não são visualmente distinguíveis entre si, caracterizando grupos *homofemas*. As transições entre visemas no sistema *MikeTalk* são realizadas de maneira linear e os efeitos da coarticulação não são contemplados pela síntese.

Em *MikeTalk*, o mapa de correspondência entre visemas é determinado através de fluxo óptico (*optical flow*), técnica de visão computacional que visa determinar e medir o movimento de objetos entre imagens (HORN; SCHUNCK, 1980). O grande diferencial na utilização da técnica de fluxo óptico, reside no fato de que cada pixel da imagem é considerado um ponto-âncora. Os autores justificam a aplicação desta técnica como uma alternativa ao processo de rotulação manual de imagens extraídas do corpus. No entanto, sua aplicação, além de computacionalmente custosa, acarreta o aparecimento de "lacunas" e ruído do tipo sal-e-pimenta (*salt-and-pepper*), caracterizado pelo aparecimento de pequenos pontos brancos e pretos nas imagens de transição, exigindo a aplicação de algoritmos adicionais para garantia da qualidade visual dos quadros da animação.

As mesmas abordagens de síntese visual da fala e metamorfose adotadas em *MikeTalk* são também empregadas em sistemas como (GOYAL; KAPOOR; KALRA, 2000) e (FARUQUIE et al., 2001). Estes sistemas, porém, adicionam ao processo de síntese a modelagem de movimentos de comunicação não-verbal (tais como expressões de tristeza e felicidade ou movimentação da cabeça e piscar de olhos) visando obter maior naturalidade da *talking head* na animação final.

Uma outra abordagem de síntese baseada em metamorfose entre visemas é apresentada em (EDGE; MADDOCK, 2003). Este sistema adota uma base de imagens com 40 visemas e contempla os efeitos da coarticulação através de modelagem baseada no estudo de Cohen e Massaro (1993). Tal modelagem é implementada através de uma função de transição temporal entre visemas utilizando funções exponenciais de influência. Um grande incoveniente do modelo teórico de Cohen e Massaro é a determinação dos parâmetros das funções de influência e a definição dos alvos articulatórios de visemas puros sem o efeito da coarticulação.

Em (EDGE; MADDOCK, 2003), é implementado um algoritmo de metamorfose guiado por pontos-âncora, utilizando funções de base radial (RBF - *Radial Basis Function*) para a operação de *warping*. Para a determinação dos pontos-âncora, os autores optaram por realizar uma análise de componentes principais (PCA - *Principal Component Analysis*), visando reduzir o número de pontos-âncora a

serem considerados durante a metamorfose. A análise PCA pressupõe um número significativo de pontos-âncora iniciais, caracterizando um processamento inerentemente custoso. Além disso, faz-se necessário definir o número de parâmetros (componentes principais) que serão utilizados na metamorfose. Como apontados pelos próprios autores, a escolha da quantidade de parâmetros não é uma questão controversa já que ela impacta a qualidade final da metamorfose. Adicionalmente, os parâmetros-âncora não necessariamente permitem uma interpretação geométrica que seria desejável para a adoção do modelo articulatório de Cohen e Massaro (1993).

Os bons resultados obtidos através da animação baseada na metamorfose entre visemas deram um grande impulso à exploração de técnicas de animação facial 2D. Tais resultados são obtidos a partir de uma base de imagens reduzida e a utilização de algoritmos de processamento de imagens bastante disseminados. Estas características são especialmente vantajosas quando confrontadas com as técnicas de animação facial 3D, onde o foto-realismo é aproximado por um elaborado, detalhado e complexo processo de modelagem da geometria tridimensional e da aparência da face.

2.4.2 Síntese concatenativa

Enquanto a abordagem baseada em metamorfose entre visemas tem suas origens nas abordagens convencionais de animação por interpolação entre poses-chave (*key-framing*), a técnica de síntese concatenativa inspira-se no paradigma utilizado por sistemas de síntese concatenativa de áudio, que geram fala sintética através da concatenação de fragmentos de fala natural (NG, 1998).

Na abordagem da síntese concatenativa, a base de imagens armazena pequenos fragmentos de vídeo correspondentes à locução de trechos de fala. Em geral, tais fragmentos compreendem a locução de dois fones (difones) ou três fones (trifones) em sequência.

O principal apelo deste tipo de abordagem resume-se no fato que a utilização de uma base de dados composta de fragmentos de vídeo capturados de uma movimentação real, permite reproduzir, com elevado nível de vídeo-realismo, a dinâmica dos movimentos articulatórios da fala, incluindo os efeitos da coarticulação. O problema associado a esta abordagem é o tamanho da base de dados utilizada, que está associada à quantidade e duração dos fragmentos de vídeo armazenados.

Um trabalho representativo desta técnica é conhecido como *Video Rewrite* (BREGLER; COVELL; SLANEY, 1997). Neste trabalho, o sistema faz uso de uma seqüencia de vídeo já existente para criar um novo vídeo com a mesma face e uma nova faixa de aúdio, num processo análogo ao de "reescrita" do vídeo original.

Assim, o corpus audiovisual é caracterizado por um trecho de vídeo já existente, sem qualquer restrição de cenário ou conteúdo de locução. A faixa de vídeo é segmentada em trifones a partir da transcrição fonética do áudio. A cada fragmento de vídeo são associadas informações sobre a posição da boca e formato dos lábios para cada quadro de produção do trifone. Os trifones e suas informações associadas, concluem a etapa de análise e constituem uma base de fragmentos de vídeo. A síntese de um novo vídeo é construída através da concatenação e apropriado ajuste de fragmentos de trifones da base.

O vídeo-realismo em *Video Rewrite* é obtido pela utilização de contextos de trifones que contemplam os efeitos da coarticulação. Porém, a qualidade final da animação pode ser limitada pelo número de contextos de trifones existentes no vídeo original. Além disso, o armazenamento de fragmentos de vídeo para um número significante de contextos apresenta elevado custo de memória e implica no armazenamento de grande quantidade de informação redundante.

Em (COSATTO; GRAF; HUANG,), (HUANG; COSATTO; GRAF, 2002), a técnica de síntese concatenativa é implementada procurando reduzir o tamanho da base de dados necessária e a redundância das informações armazenadas. A partir de um corpus audiovisual capturado em condições controladas, são extraídas imagens rotuladas de acordo com a transcrição fonética temporizada da faixa de áudio. Levando-se em consideração o grande número de imagens obtidas, o processo de construção da base de imagens neste sistema inclui a aplicação da análise PCA (*Principal Component Analysis*) para redução do tamanho da base de imagens.

A síntese concatenativa adotada neste sistema é implementada através de um algoritmo de seleção de visemas, em que a unidade básica de seleção é caracterizada por trifones. Para cada fone de uma nova animação a ser sintetizada, o sistema busca na base de imagens sequências de visemas que melhor caracterizam visualmente e/ou foneticamente a produção do trifone que tem como fone central o fone alvo. A animação final é obtida através da concatenação das sequências resultantes do processo de seleção de visemas da base de imagens.

Para atacar o problema da base de imagens com tamanho inerentemente excessivo, modificações da implementação da técnica têm sido propostas. Em (EDGE M. SANCHES, 2004), por exemplo, os autores partem de um corpus audiovisual em que o conteúdo de locução é definido com base em uma aplicação específica (como datas e horários, por exemplo). No momento da síntese, procura-se maximizar a coincidência de contextos fonéticos incluindo difones, sílabas, palavras ou fragmentos de texto, tal como eles foram gravados pelo corpus. Já (KSHIRSAGAR; MAGNENAT-THALMANN, 2003) adota uma base de visílabas (*visyllables*) definidas nesse trabalho, como os correspondentes visuais da produção acústica de sílabas.

Tendo-se em perspectiva as diversas implementações apresentadas, é possível destacar os seguintes inconvenientes da abordagem baseada em síntese concatenativa:

- o número de imagens armazenadas na base de imagens é significativamente maior que o número de imagens de sistemas baseados na metamorfose entre visemas;

- a concatenação de fragmentos de vídeo na maioria das vezes não exclui a aplicação de algoritmos de processamento de imagens que garantam uma transição suave entre as fronteiras dos fragmentos.

2.4.3 Trajetória de visemas em espaço multidimensional

O surgimento de sistemas computacionais com capacidade de processamento e armazenamento de dados cada vez maiores viabilizou a implementação de sistemas de animação facial 2D baseados em elaborados e complexos algoritmos de análise e processamento das imagens e fragmentos de vídeo de um abrangente corpus audiovisual.

Uma das características compartilhadas por sistemas deste tipo é que a totalidade dos milhares de quadros resultantes da captura do corpus audiovisual é analisada e processada durante o processo de construção da base de imagens (COSATTO; GRAF, 2000), (EZZAT; GEIGER; POGGIO, 2002), (BASU et al.,). De maneira bastante ilustrativa, os autores de (EZZAT; GEIGER; POGGIO, 2002) afirmam que, enquanto o processo de captura do corpus audiovisual de seu sistema dura apenas 15 minutos, a execução dos algoritmos de processamento automático das imagens capturadas leva vários dias até ser concluída.

O resultado deste processo é a constituição de uma base de imagens organizada em um espaço multidimensional, no qual as imagens são indexadas através de parâmetros específicos relacionados às suas características visuais. Nestes sistemas, a síntese da animação consiste na definição de uma trajetória dentro deste espaço multidimensional de visemas.

Um exemplo da aplicação deste tipo de técnica é descrito no trabalho (COSATTO; GRAF, 2000). As imagens extraídas do corpus audiovisual, capturado em condições controladas, passam por uma sofisticada cadeia de processos visando a detecção e extração de elementos da face e suas características (GRAF; COSATTO; EZZAT, 2000). Nesta abordagem, a região de olhos, sobrancelhas e lábios são extraídas das imagens e formam bases de imagens independentes. Com isso, este sistema é capaz de reproduzir não apenas a movimentação articulatória da fala, mas também modela a prosódia visual através da síntese de movimentação das sobrancelhas, piscar de olhos e cabeça (GRAF et al., 2002).

As imagens correspondentes aos visemas são rotuladas com parâmetros como: contexto fonético em que o visema foi produzido, pose da cabeça, número do quadro correspondente na sequência de vídeo original, entre outros. Além disso, através de algoritmos de visão computacional, em uma de suas implementações o sistema organiza estes visemas num espaço tridimensional de acordo com as informações visuais de: largura da boca (distância entre os cantos dos lábios), posição do lábio superior e posição do lábio inferior (vide Figura 2.7).

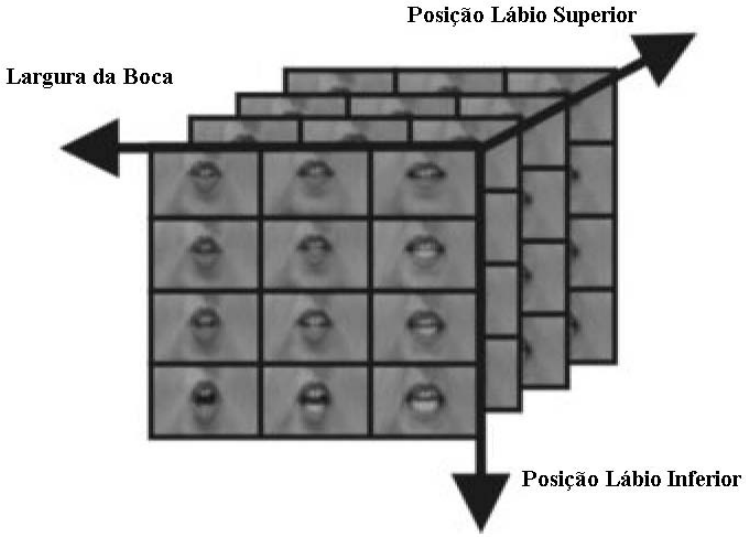


Fig. 2.7: Espaço tridimensional de imagens (adaptado de (COSATTO; GRAF, 2000))

A partir da transcrição fonética temporizada, o sistema busca na base de imagens os visemas cujos contextos fonéticos são mais próximos do contexto alvo a ser sintetizado. Este processo de busca resulta na construção um grafo para a animação que contém uma lista de visemas candidatos para cada quadro da animação final. Este grafo permite calcular funções de custo que determinam o custo de transição entre candidatos de quadros consecutivos. Uma vez que os custos entre nós do grafo são determinados, o caminho de custo mínimo é determinado utilizando-se um algoritmo de Viterbi (FORNEY, 1973). A Figura 2.8 mostra o grafo utilizado e o caminho mínimo para uma determinada animação representado pelas flechas mais escuras.

A modelagem da coarticulação é abordada de duas maneiras na implementação deste sistema.

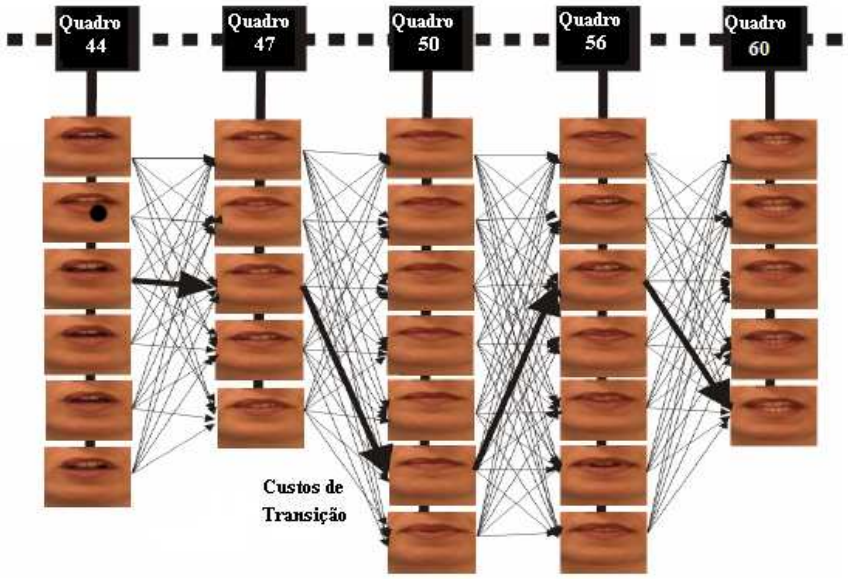


Fig. 2.8: Visemas candidatos, transições de custo e caminho mínimo estabelecido pelo algoritmo de Viterbi (adaptado de (COSATTO; GRAF, 2000))

Primeiramente, a busca de visemas candidatos na base de imagens leva em consideração uma função baseada na proposta de (COHEN; MASSARO, 1993) que ajuda a classificar quais visemas possuem contexto fonético mais próximo do contexto alvo a ser sintetizado. Em segundo lugar, é possível afirmar que para os casos em que o alvo a ser sintetizado possui características idênticas a um determinado contexto previamente capturado pelo corpus audiovisual, o algoritmo implementado irá reaproveitar quadros capturados em sequência pelo corpus original, e assim, obtém-se a melhor situação possível de reprodução dos movimentos articulatórios. A Figura 2.9 mostra a trajetória descrita no espaço multidimensional de visemas para uma determinada sequência de animação.

Uma outra abordagem interessante é mostrada em (EZZAT; GEIGER; POGGIO, 2002). Neste sistema, a primeira etapa de processamento das imagens do corpus audiovisual envolve a aplicação de algoritmos que visam identificar redundância nas características visuais do universo de imagens existentes, resultando na criação de um espaço bidimensional de 46 imagens organizadas segundo um modelo multidimensional de metamorfose (*Multidimensional Morphable Model*), ou MMM. Neste espaço bidimensional, as imagens são indexadas por valores que representam informações sobre a **aparência** e **forma** das imagens, ou ainda, podem ser traduzidas em características de textura e geométricas das faces. Uma vez definido o MMM, é possível gerar novas imagens com posturas labiais que não estão presentes na base, a partir do fornecimento de novos valores para os parâmetros de entrada do modelo, num processo de síntese de novas imagens.

Além da síntese, o modelo estabelecido é também utilizado para projetar as imagens capturadas do corpus audiovisual neste espaço multidimensional, retornando-se uma trajetória de parâmetros de **aparência** e **forma** para cada conteúdo pronunciado pelo locutor. Através da análise destas trajetórias, o sistema é treinado, derivando-se um modelo de coarticulação que não é genérico, mas que reflete as características da locução capturada pelo corpus audiovisual. A partir de um novo conteúdo a ser sintetizado, o sistema é capaz de ponderar as informações de velocidade do discurso na nova

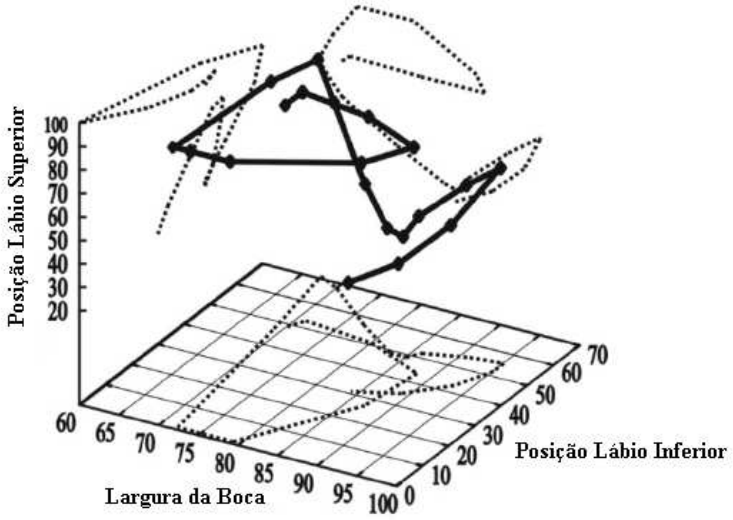


Fig. 2.9: Os visemas selecionados da base de imagens determinam uma trajetória no espaço multidimensional de visemas (adaptado de (COSATTO; GRAF, 2000))

locução com as informações de coarticulação provenientes da etapa de treinamento, gerando uma nova trajetória no espaço multidimensional de imagens.

A síntese baseada na trajetória de visemas em um espaço multidimensional tipicamente apresenta resultados convincentes na reprodução da movimentação articulatória visível da fala, caracterizando uma solução apropriada para aplicações em que o alto nível de vídeo-realismo seja um pré-requisito. No entanto, tais resultados são consequência direta dos elaborados modelos e etapas de processamento envolvidos, e pelo intenso reaproveitamento de informações extraídas do corpus audiovisual. Por este motivo, tais sistemas são computacionalmente custosos tanto na etapa de construção da base de imagens quanto em seu processo de síntese. Esta característica representa uma grande desvantagem na implementação de aplicações que visam a geração da animação em tempo real ou aplicações a serem embarcadas em dispositivos de pequena capacidade.

2.5 Comentários Finais

Desde o nosso nascimento somos treinados na complexa e detalhada estrutura da face e no reconhecimento de suas pistas visuais durante a fala. Por este motivo, possuímos acurado e exigente senso crítico na apreciação de animações faciais por computador, fato que tem motivado um grande número de pesquisas voltadas à animação vídeo-realista sincronizada com a fala.

A geração de uma animação facial em sincronia e harmonia com a fala exige a reprodução convincente e realista dos movimentos articulatórios associados à realização dos vários segmentos sonoros da língua. Para tanto, além da identificação das posturas características dos gestos articulatórios associados aos segmentos sonoros, faz-se necessária a representação das transições entre estas posturas considerando os efeitos da coarticulação. Os efeitos da coarticulação se manifestam pela alteração do padrão articulatório de um determinado segmento sonoro pela influência da articulação de outro

adjacente ou, e em menor grau, próximo na cadeia da produção sonora.

Neste capítulo foram identificadas duas abordagens principais para a animação facial por computador: a animação baseada na manipulação de modelo geométrico, ou 3D, e a animação baseada em imagens, ou 2D. A característica inerentemente realista das fotografias, empresta às animações 2D um grau de realismo aproximado apenas através de grandes esforços pela animação 3D, que tipicamente apresenta uma marcante aparência artificial.

Considerando-se o contexto 2D, neste capítulo foram introduzidos os principais conceitos relacionados a sistemas de animação facial por meio de um recorte da literatura pertinente, apresentando-se o estado da arte para sistemas deste tipo.

A Tabela 2.1 sintetiza o panorama histórico dos principais trabalhos citados e ressalta as principais diferenças observadas entre as diferentes implementações apresentadas a partir da análise de aspectos como: a origem do corpus audiovisual, a natureza e tamanho da base de imagens implementada pelo sistema, características da modelagem da movimentação articulatória visível da fala, estratégia de síntese adotada e a modelagem de movimentação relacionada à comunicação não-verbal.

Tab. 2.1: Comparação entre sistemas de animação facial 2D

Trabalho	**Corpus Audiovisual**	**Natureza da Base de Imagens**	**Tamanho da Base de Imagens**	**Modelagem Movimentação Articulatória**	**Estratégia de Síntese**	**Modelagem Comunicação Não-Verbal**
Scott et al., 1994	Captura controlada.	Imagens Faciais	Reduzida: 50 Visemas	Não modela coarticulação.	Metamorfose entre Visemas	Não
Bregler; Covell; Slaney, 1997	Trecho de vídeo sem restrições.	Fragmentos de vídeo (trifones).	Extensa	Modela coarticulação através da concatenação de fragmentos de vídeo correspondentes a trifones.	Síntese concatenativa.	Não
Ezzat; Poggio, 1998	Captura controlada.	Imagens Faciais	Reduzida: 16 Visemas	Não modela coarticulação.	Metamorfose entre Visemas	Não
Cosatto; Graf, 2000	Captura controlada.	Sequência de quadros extraídos do corpus. Armazena separadamente imagens da região dos lábios, olhos e sobrancelhas.	Extensa	Modela coarticulação através de algoritmos que permitem selecionar quadros da base de imagens com base no contexto fonético alvo a ser sintetizado e na obtenção de transições suaves da animação.	Trajetória de Visemas em Espaço Multidimensional	Inclui movimentação da cabeça com o auxílio de um modelo geométrico 3D. Animação reproduz prosódia visual.
Huang; Cosatto; Graf, 2002	Captura controlada.	Imagens faciais correspondentes a sequência de quadros extraídos do corpus.	Extensa	Modela coarticulação através concatenação de quadros capturados em contextos fonéticos (trifones) semelhantes ao alvo a ser sintetizado.	Síntese concatenativa	Não
Ezzat; Geiger; Poggio, 2002	Captura controlada.	Imagens faciais.	Reduzida: 46 imagens	Modela coarticulação através de um modelo multidimensional de metamorfose.	Trajetória de Visemas em Espaço Multidimensional	Não
Edge; Maddock, 2003	Captura controlada.	Imagens Faciais	Reduzida: 40 visemas	Contempla os efeitos da coarticulação através de modelagem baseada no estudo de Cohen e Massaro (1993).	Metamorfose entre visemas.	Não

É possível afirmar que, no princípio, os convincentes resultados obtidos através de sistemas de animação facial 2D caracterizaram um atalho na obtenção de animações vídeo-realistas, que apesar de menos flexíveis quando comparadas a abordagem 3D, necessitavam de menos recursos computacionais e utilizavam modelos mais simplificados. No entanto, algumas implementações recentes aplicam técnicas de síntese de animação facial 2D que se assemelham em grau de complexidade e elaboração aos sistemas baseados na manipulação de modelos geométricos. De fato, a combinação das duas abordagens vem sendo explorada como uma alternativa emergente e promissora na obtenção de *talking heads* extremamente flexíveis em sua manipulação e com elevado grau de vídeo-realismo.

No entanto, as técnicas e modelagens mais simples de sistemas 2D não deixarão de ter seu espaço. Esta afirmação se justifica pela crescente disponibilidade de múltiplos meios de acesso à informação, fazendo surgir inúmeras aplicações que impulsionam o desenvolvimento de sistemas de *talking heads*. Nesta nova realidade, a tecnologia de animação facial deverá ser embarcada em dispositivos pequenos

e portáteis que, quando comparados a sistemas *desktop*, geralmente apresentam menor abundância de recursos de processamento e memória.

É neste contexto contemporâneo e com foco nesta realidade que este trabalho se desenvolve.

O presente trabalho apresenta a implementação de um sistema de animação facial 2D, enfatizando a reprodução realista da movimentação articulatória visível para o Português do Brasil e a implementação de um processo de síntese computacionalmente viável para execução em dispositivos com recursos limitados de processamento e memória. Considerando-se a abordagem de síntese adotada, o sistema implementado revisita a técnica de metamorfose entre visemas apresentando sua contribuição através da adoção de uma modelagem de animação visual da fala que contempla os efeitos da coarticulação a partir de uma base de imagens de apenas 34 visemas.

O gráfico da Figura 2.10 apresenta uma comparação entre os principais trabalhos citados neste capítulo, visando situar este trabalho no universo de sistemas de animação facial 2D e ressaltar a visão de desenvolvimento adotada. Com este propósito, os trabalhos foram organizados segundo critérios de flexibilidade e realismo, tendo-se como base os vídeos disponibilizados por seus autores e analisando-se suas principais características de implementação. No contexto do gráfico apresentado, o critério de flexibilidade engloba aspectos da implementação tais como: particularidades do processo de captura do corpus audiovisual, tamanho da base de imagens, complexidade dos algoritmos de pré-processamento e de busca e seleção das imagens da base e custo computacional da estratégia de síntese adotada. Por outro lado, o critério de realismo está relacionado ao grau de fidelidade alcançado na reprodução dos aspectos visuais da face, bem como da dinâmica articulatória da fala, incluindo a prosódia visual, movimentos fisiológicos e expressão de emoções.

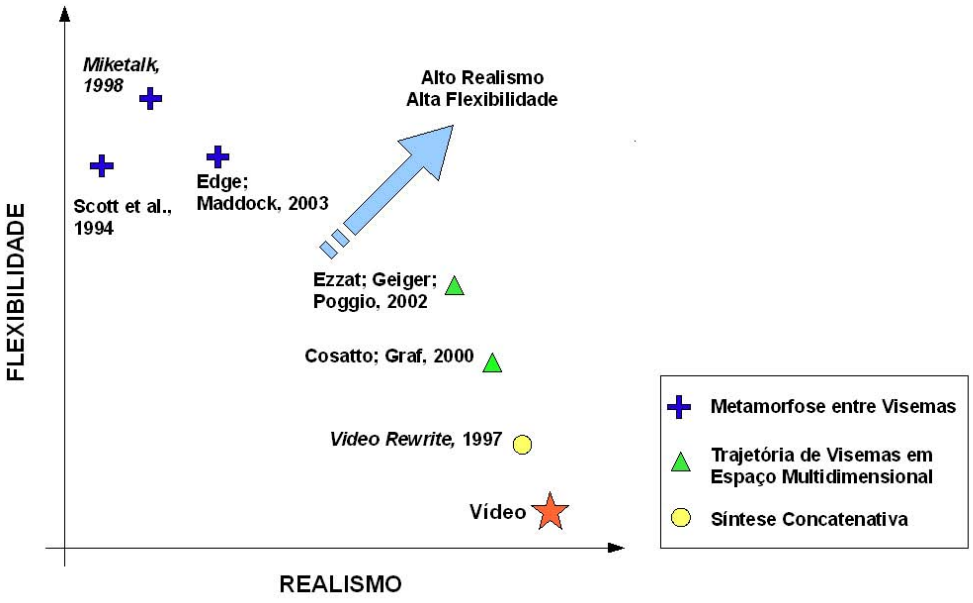


Fig. 2.10: Comparação entre sistemas de animação facial 2D segundo critérios de flexibilidade e realismo

A partir destes critérios, o vídeo de uma face real é apresentado como a solução de maior realismo possível mas que apresenta o menor nível de flexibilidade, uma vez que cada novo conteúdo de locu-

ção exige um novo arranjo de captura, com elevado custo de produção e armazenagem em memória dos conteúdos gerados. No extremo oposto, os sistemas de animação baseados na metamorfose entre visemas apresentam as soluções de maior flexibilidade e menor nível de realismo, adotando bases de imagens extremamente reduzidas e modelos articulatórios bastante simplificados.

Seguindo esta análise, o presente trabalho apresenta uma estratégia de síntese baseada na metamorfose entre visemas, com reduzido tamanho de base de imagens e características de implementação que mantêm o alto nível de flexibilidade característicos desta abordagem. No entanto, a implementação proposta apresenta uma contribuição ao gerar animações com maior nível de realismo devido, principalmente, à adoção de uma modelagem articulatória mais acurada. Desta maneira, este trabalho representa um passo na direção do desenvolvimento de técnicas que permitam a implementação de sistemas com alto nível de flexibilidade e elevado nível de vídeo-realismo, que possam ser utilizados nos diversos contextos possíveis em que interfaces baseadas na comunicação face-a-face sejam demandadas.

Os capítulos que se seguem focam na apresentação da metodologia empregada para construção da base de imagens e na descrição do processo de síntese implementado pelo sistema de animação facial 2D implementado neste trabalho.

Capítulo 3

Construção da Base de Imagens

3.1 Introdução

Como discutido no Capítulo 2, a modelagem da face em sistemas de animação facial 2D é implementada através da construção de uma base de imagens extraídas de um corpus audiovisual. A construção da base de imagens envolve importantes aspectos que afetam diretamente o nível de vídeo-realismo observado nas animações geradas pelo sistema e que refletem a estratégia de síntese adotada.

A qualidade visual das imagens capturadas pelo corpus audiovisual e os algoritmos de processamento utilizados para a preparação das imagens da base, por exemplo, são fatores que influenciam diretamente o foto-realismo dos quadros sintetizados da animação.

Por outro lado, o número de imagens armazenadas e a maneira como elas são organizadas e indexadas na base refletem a abordagem adotada para a reprodução da movimentação articulatória vísivel e também definem importantes características de complexidade, flexibilidade e aplicabilidade do sistema de animação facial 2D, conforme discutido no Capítulo 2.

Outro aspecto importante é a natureza das imagens armazenadas. É possível, por exemplo, armazenar imagens da face como um todo (EZZAT; POGGIO, 1998), diferentes ângulos ou poses da cabeça (GOYAL; KAPOOR; KALRA, 2000), expressões faciais (FARUQUIE et al., 2001) ou imagens de regiões limitadas da face como lábios, olhos e sobrancelhas (COSATTO; GRAF, 2000). Tais abordagens conferem diferentes graus de liberdade à síntese da animação facial e possibilitam a reprodução, ou não, de movimentações que não estão diretamentamente associadas à produção da fala mas que atribuem maior naturalidade à *talking head*.

Este capítulo descreve a metodologia utilizada para a criação da base de imagens do sistema de animação facial 2D apresentado neste trabalho. A construção da base de imagens seguiu a cadeia de processos apresentada na Figura 3.1.

Este capítulo aborda inicialmente, na Seção 3.2, o conceito de visemas dependentes de contexto fonético. Considerando a estratégia de síntese adotada, baseada na técnica de metamorfose entre visemas, a adoção da modelagem da movimentação articulatória através da utilização de visemas dependentes de contexto representa um importante aspecto da abordagem proposta neste trabalho. Mais especificamente, a base de imagens implementada é constituída por um conjunto de visemas dependentes de contexto para o Português do Brasil.

A partir da definição de visemas dependentes de contexto, é possível descrever o processo empregado para captura do corpus audiovisual. O corpus audiovisual foi gerado a partir da captura em

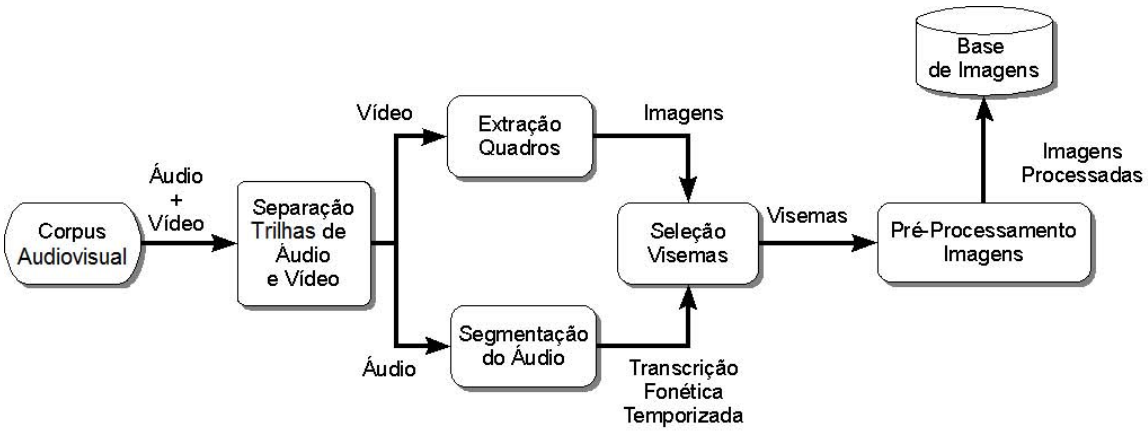


Fig. 3.1: Construção da Base de Imagens

áudio e vídeo, em condições controladas, da face de uma locutora feminina pronunciando conteúdos previamente estabelecidos (Seção 3.3).

Conforme mostrado na Figura 3.1, as trilhas de vídeo e áudio resultantes do processo de captura são dessincronizadas e separadas.

Na Seção 3.4, descreve-se o processo de transcrição fonética temporizada da trilha de áudio correspondente, que permitiu associar as imagens capturadas à produção dos diferentes fonemas da língua.

A partir desta informação, foi possível selecionar e extrair quadros individuais da trilha de vídeo. Estes quadros foram pré-processados visando a uniformização de enquadramento, formatação e construção da base de imagens.

Neste trabalho, a base de imagens é caracterizada por 34 visemas ou imagens fotográficas representando diferentes posturas labiais estáticas. O número reduzido de imagens armazenadas é reflexo da adoção da técnica de síntese baseada na metamorfose entre visemas e da modelagem da movimentação articulatória visível por meio de visemas dependentes de contexto. Os critérios utilizados para a seleção das imagens e os algoritmos de processamento das imagens são descritos na Seção 3.5.

A etapa de modelagem da face é finalizada através da consolidação das informações extraídas da transcrição fonética temporizada do áudio e de parâmetros visuais das imagens selecionadas, permitindo a construção de uma base de visemas devidamente rotulados com informações auxiliares à etapa de síntese. A descrição completa da base de imagens é apresentada na Seção 3.6.

Conforme discutido anteriormente, os diferentes aspectos envolvidos na construção da base de imagens influenciam diretamente o nível de vídeo-realismo das animações geradas por um sistema de animação facial 2D. Assim, a Seção 3.7 encerra o capítulo destacando as principais características da base de imagens descrita, ressaltando suas principais características e respectivos impactos no vídeo-realismo das animações geradas pelo sistema.

3.2 Visemas Dependentes do Contexto Fonético para o Português do Brasil

Os diferentes sons da fala são produzidos pela modificação controlada do fluxo de ar pulmonar em sua passagem pelo trato vocal, associada a um posicionamento e movimentação característicos de elementos articuladores tais como lábios, palato, cavidade nasal e cordas vocais.

Considerando-se o problema de reproduzir sinteticamente a movimentação articulatória, é importante notar que grande parte desta movimentação ocorre no interior da cavidade oral sem que seja possível uma fácil visualização. Consequentemente, é possível afirmar que a percepção visual para discriminação de diferentes segmentos sonoros é menos eficaz que a percepção auditiva, o que torna possível associar visemas a agrupamentos de segmentos sonoros não distinguíveis visualmente. Tais agrupamentos são denominados homofemas (DE MARTINO, 2005).

A identificação de grupos homofemas e seus respectivos visemas para uma determinada língua caracteriza uma importante observação que permite a simplificação e redução da base de imagens em sistemas de animação facial 2D. Este abordagem é adotada, por exemplo, em (EZZAT; POGGIO, 1998) que utiliza apenas 16 visemas para representar aproximadamente 50 fonemas da língua inglesa. Tal recurso é também utilizado por sistemas que elegem, dentre várias imagens candidatas, visemas com características visuais mais apropriadas a um determinado quadro da animação. A existência de homofemas, neste caso, permite que se considere um maior número de imagens candidatas correspondentes a visemas associados a um mesmo grupo homofema (COSATTO; GRAF, 2000).

Considerando-se o Português do Brasil, DE MARTINO (2005) realizou trabalho pioneiro ao identificar os visemas característicos desta língua e organizá-los em grupos homofemas. As tabelas 3.1 e 3.2 apresentam os grupos homofemas identificados em (DE MARTINO, 2005) para fones consonantais e vocálicos, respectivamente. Na Tabela 3.1, por exemplo, a primeira linha contempla o grupo homofema [p,b,m], que é simplificadamente representado pelo fone [p]. Neste trabalho, a representação textual dos segmentos da fala é realizada utilizando o Alfabeto Fonético Internacional (IPA - *International Phonetic Alphabet*) (INTERNATIONAL PHONETIC ASSOCIATION, 1999). A representação dos fones adota a convenção da Associação Fonética Internacional em que são utilizados símbolos IPA entre colchetes, como por exemplo [p].

Além da identificação de visemas para o Português do Brasil, DE MARTINO (2005) também identificou variações da movimentação articulatória visível para visemas pertencentes a um mesmo grupo homofema produzidos porém, em diferentes contextos fonéticos. Tais variações observadas nos padrões articulatórios são manifestações dos efeitos da coarticulação pela influência dos fones vizinhos representados pelos contextos fonéticos considerados nesse trabalho. Os visemas identificados em seu trabalho são denominados visemas dependentes de contexto.

Neste trabalho, os visemas dependentes de contexto identificados para o Português do Brasil foram utilizados como referência para a representação visual dos sons da fala para o sistema implementado. Ao todo, foram considerados 34 visemas, correspondendo a: 22 visemas consonantais, 11 visemas vocálicos e 1 visema representante da posição de repouso, ou silêncio.

A Tabela 3.3 apresenta os visemas consonantais adotados. A primeira coluna da tabela apresenta o grupo homofema considerado. A segunda coluna apresenta a simbologia adotada para a representação dos visemas. Neste trabalho, os visemas são representados por símbolos IPA entre "<" e ">". Os vários visemas associados a contextos diferentes de um mesmo grupo homofema são identificados

Tab. 3.1: Homofemas consonantais e fones representantes (extraído de (DE MARTINO, 2005)).

Grupo Homofema	Fone Representante
[p,b,m]	[p]
[f,v]	[f]
[t,d,n]	[t]
[s,z]	[s]
[l]	[l]
[k,g]	[k]
[ʃ, ʒ]	[ʃ]
[ʎ, ɲ]	[ʎ]
[ɾ],[ɣ]	[ɾ]

Tab. 3.2: Homofemas vocálicos e fones representantes (extraído de (DE MARTINO, 2005))

Grupo Homofema	Fone Representante
[i,ĩ]	[i]
[a,ɛ,ɐ̃]	[a]
[u,o,õ,ũ]	[u]
[ɪ]	[ɪ]
[ɐ,e,ɔ,ẽ]	[ɐ]
[ʊ]	[ʊ]

por um índice numérico. Na terceira coluna têm-se os contextos fonéticos associados a cada visema.

Tab. 3.3: Visemas consonantais dependentes de contexto (extraído de (DE MARTINO, 2005))

Grupo Homofema	Visemas	Contextos Fonéticos
[p,b,m]	$< p_1 >$	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]
	$< p_2 >$	[pu] [upɪ] [upʊ]
[f,v]	$< f_1 >$	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]
	$< f_2 >$	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]
[t,d,n]	$< t_1 >$	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]
	$< t_2 >$	[ta] [atɐ]
[s,z]	$< s_1 >$	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]
	$< s_2 >$	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]
[l]	$< l_1 >$	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]
	$< l_2 >$	[la] [ilɐ] [alɪ] [alɐ]
	$< l_3 >$	[lu]
	$< l_4 >$	[ilʊ] [ulʊ]
[ʃ, ʒ]	$< ʃ_1 >$	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]
	$< ʃ_2 >$	[ʃu] [uʃʊ]
[ʎ, ɲ]	$< ʎ_1 >$	[ʎi] [ʎa] [iʎɪ] [iʎɐ] [aʎɪ] [aʎɐ]
	$< ʎ_2 >$	[ʎu] [uʎɪ] [uʎɐ]
	$< ʎ_3 >$	[iʎʊ] [aʎʊ] [uʎʊ]
[k,g]	$< k_1 >$	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]
	$< k_2 >$	[ka] [akɐ]
	$< k_3 >$	[ku] [ikʊ] [akʊ] [ukʊ]
[ɾ],[ɣ]	$< ɾ_1 >$	[ɣi] [ɣa] [iɾɪ] [iɾɐ] [aɾɪ] [aɾɐ] [uɾɐ]
	$< ɾ_2 >$	[ɣʊ] [iɾʊ] [aɾʊ] [uɾɪ] [uɾʊ]

Observando-se a coluna "Contextos Fonéticos" da tabela, é possível notar que os contextos fonéticos associados a cada visema não contemplam a totalidade de contextos fonéticos possíveis na língua. Os contextos fonéticos contemplados consideram a coarticulação adjacente em dois tipos de contexto:

- $\#CV$ - segmento consonantal (C) entre silêncio (#) e segmento vocálico (V);
- V_1CV_2 - segmento consonantal (C) entre dois segmentos vocálicos (V_1 e V_2).

onde (DE MARTINO, 2005):

- C={[p,f,t,s,l,k,ʃ,ʎ,l,(ɣ)ɾ]};
- $V, V_1 \in$ {[i,a,u]};

Tab. 3.4: Visemas vocálicos dependentes de contexto (extraído de (DE MARTINO, 2005))

Grupo Homofema	Visemas	Contextos Fonéticos
[i,ĩ]	$< i_1 >$	Todos os contextos exceto [tit] e [ʃiʃ].
	$< i_2 >$	[tit] e [ʃiʃ].
[e,ẽ]	$< e >$	Todos os contextos.
[ɛ]	$< \varepsilon >$	Todos os contextos.
[a,ɐ̃]	$< a >$	Todos os contextos.
[ɔ]	$< \text{ɔ} >$	Todos os contextos.
[o,õ]	$< o >$	Todos os contextos.
[u,ũ]	$< u >$	Todos os contextos.
[ɪ]	$< \text{ɪ} >$	Todos os contextos.
[ɐ]	$< \text{ɐ} >$	Todos os contextos.
[ʊ]	$< \text{ʊ} >$	Todos os contextos.

- $V_2 \in \{[\text{ɪ,ɐ,ʊ}]\}$.

Considerando-se o conjunto de fones possíveis para o segmento consonantal *C* deve-se notar que o som correspondente ao fone [ɾ] (produzido, por exemplo, ao se pronunciar o "R" da palavra "PARA"), designado "alveolar tepe", não ocorre no início de palavras na língua portuguesa. Por este motivo o fone [ɾ] é considerado nos contextos $V_1\mathbf{C}V_2$, enquanto que nos contextos $\#CV$ foi considerado o fone [ɣ]. Os outros fones pertencentes ao conjunto *C* compreendem os fones representantes de cada grupo homofema tais como apresentados na Tabela 3.1. Assim, por exemplo, considerando-se o grupo homofema [p,b,m], o segmento consonantal representante utilizado é [p].

De maneira similar, os segmentos vocálicos V_1 e V_2 também representam grupos homofemas vocálicos conforme apresentado na Tabela 3.2.

A Tabela 3.4, por sua vez, apresenta os visemas vocálicos adotados neste trabalho. Observa-se que o grupo homofema [i,ĩ] é o único que contém mais de um visema para representar diferentes contextos fonéticos. Nesta tabela, os contextos fonéticos contemplados ($\#CV$ ou V_1CV_2) possuem a mesma definição e tratativa descritas para a Tabela 3.3.

As tabelas 3.3 e 3.4 representam o essência da modelagem movimentação articulatória adotada neste trabalho. A adoção de visemas dependentes de contexto para a síntese baseada na metamorfose entre visemas, permite contemplar os efeitos da coarticulação na modelagem da movimentação articulatória visível da fala. Durante o processo de síntese, todos os contextos fonéticos que ocorrem são mapeados para um dos dois tipos de contexto ($\#CV$ ou V_1CV_2), através de tabela de mapeamento mostrada na Seção 4.3.1 do Capítulo 4. Conforme discutido no Capítulo 2, a modelagem da coarticulação é um aspecto essencial na obtenção de animações vídeo-realistas.

3.2.1 Corpus Linguístico

O primeiro passo na captura de imagens correspondentes aos visemas dependentes de contexto mostrados na seção anterior, consistiu na definição dos conteúdos a serem pronunciadas pela locutora durante o processo de gravação do áudio e vídeo.

Uma vez que a constituição de um corpus audiovisual é um processo custoso, que envolve uma série de recursos físicos e humanos, a definição dos conteúdos a serem pronunciados visou a coleta de um número de contextos fonéticos e situações de locução abrangentes e muitas vezes superiores aos realmente empregados na metodologia de construção da base de imagens implementada neste trabalho. Dois tipos de conteúdos foram definidos: um conjunto de 138 palavras sem significado (logatomas) e um conjunto de 27 frases que, quando consideradas em conjunto, possuem amostras de todos os fonemas da língua portuguesa. A pronúncia de frases foi realizada com o objetivo de se obter amostras de visemas em contextos fonéticos variados e ritmo de locução mais dinâmico, caracterizando um discurso mais natural do que a locução de logatomas. O Apêndice A apresenta todos os itens pronunciados pela apresentadora durante o processo de captura do corpus audiovisual.

No contexto deste trabalho, apenas o material audiovisual resultante da locução dos logatomas foi processado. No entanto, o material audiovisual resultante da locução de frases foi essencial para diversas atividades: análise visual da dinâmica da fala na produção de fones em diversos contextos fonéticos; comparação subjetiva entre quadros filmados de uma face real e quadros gerados sinteticamente pelo sistema de animação; obtenção de material para avaliação do sistema de síntese e, finalmente, a possibilidade de evolução do sistema de síntese a partir de uma expansão da base de imagens (Capítulo 7).

O conjunto de logatomas pronunciados foi definido tendo-se como objetivo a reprodução dos contextos fonéticos contemplados nas tabelas 3.3 e 3.4, mas também incluíram contextos fonéticos que caracterizam ditongos ou encontros consonantais frequentes do Português do Brasil.

Assim, os logatomas pronunciados podem ser divididos em 4 conjuntos com as seguintes regras de formação (vide Apêndice A):

- **Logatomas paroxítonos do tipo** $'CV_1CV_2$[1], com:
  - C={[p,f,t,s,l,ʃ,ʎ,k,ɾ,ɣ]};
  - V_1={[i,a,u]};
  - V_2={[ɪ,ɐ,ʊ]}.

  resultando em 90 logatomas diferentes.

- **Ditongos do tipo** V_1V_2, com:
  - V_1={[i,e,ɛ,a,ɔ,o,u]};
  - V_2={[ɪ,ɐ,ʊ]}.

  resultando em 21 logatomas diferentes.

- **Encontros consonantais do tipo** C_1C_2V, com:

[1] O símbolo IPA " ′ ", antes de C, precede e marca a sílaba acentuada.

  - C_1={[d,p,t,k,f]};
  - C_2={[ɾ]};
  - V={[ɪ,ɐ,ʊ]}.

  resultando em 15 logatomas diferentes.

- **Encontros consonantais do tipo** C_1C_2V, com:

  - C_1={[p,t,k,f]};
  - C_2={[l]};
  - V={[ɪ,ɐ,ʊ]}.

  resultando em 12 logatomas diferentes.

3.3 Captura Audiovisual

Para a identificação das imagens representando os visemas dependentes de contexto para o Português do Brasil, foram efetuadas gravações em vídeo da face de uma locutora feminina enunciando logatomas e frases (Seção 3.2.1). As gravações foram realizadas nas dependências da Fundação CPqD, utilizando-se a infra-estrutura do Laboratório da Gerência de Serviços e Aplicações Multimídia desta instituição.

A apresentadora foi gravada com auxílio de uma câmera SONY DSR-PD170, posicionada frontalmente em relação à sua face, a uma distância de aproximadamente 1 metro. A captura foi realizada no padrão NTSC (*National Television Systems Committee*), com frequência de captura de 29.97 quadros por segundo. Para a captura do áudio foi utilizado um microfone da marca AKG, modelo C414B-ULS, posicionado de maneira a não aparecer na imagem do vídeo. A apresentadora foi posicionada diante de um fundo de cor azul e a iluminação foi disposta de maneira a iluminar adequadamente sua face, tomando-se cuidado para não criar sombras no fundo ou na região do pescoço.

Os logatomas e frases foram exibidos em um monitor posicionado à frente da apresentadora, caracterizando um mecanismo de *teleprompter*. A apresentação de cada item a ser pronunciado era controlada a partir de um computador portátil executando aplicativo Java especialmente desenvolvido para esta atividade. O aplicativo foi dotado de capacidades básicas de navegação pela lista de itens a serem pronunciados, tais como: ir para o próximo, repetir o atual, voltar para a anterior, ou ir para um determinado item. Outra característica do aplicativo era a geração de um tom de 1 kHz e duração de 0,5 segundo, utilizado para marcar o início da apresentação de cada item a ser pronunciado.

O mecanismo de *teleprompter* foi operado por um membro da equipe de gravação localizado dentro do estúdio. Dentro do estúdio de gravação, além da apresentadora e do operador de *teleprompter*, estava presente também um membro responsável por acompanhar e conferir a lista de conteúdos pronunciados e monitorar a correta locução destes. Um quarto membro da equipe de gravação foi responsável por monitorar o enquadramento e expressão da face no vídeo a partir de uma sala de controle contígua ao estúdio. Um profissional técnico foi responsável pela configuração e operação inicial dos equipamentos utilizados durante para o processo de gravação.

Fig. 3.2: Imagem capturada no processo de gravação do corpus audiovisual

Ao se detectar uma falha de locução ou de enquadramento, optou-se por repetir imediatamente o item, até a sua produção correta. A apresentadora foi instruída a pronunciar cada item apenas após o tom de 1 kHz e articulá-lo a partir de uma posição de repouso, com a boca fechada e os dentes cerrados, devendo retornar a esta posição após a produção acústica. A locução foi realizada sempre de maneira neutra, sem a expressão de emoções.

A face foi gravada sem quaisquer tipo de marcadores visuais. Além disso, a apresentadora não utilizou quaisquer acessórios, como colar ou brincos. Seus cabelos foram presos e sua cor de roupa foi escolhida de modo a ser visualmente contrastante com a cor de sua pele e do fundo do cenário. A Figura 3.2 apresenta um quadro extraído do vídeo resultante do processo de filmagem.

O material capturado totalizou aproximadamente 40 minutos de gravação. O conteúdo foi originalmente gravado em formato digital Mini-DV. Através de um gravador e reprodutor de vídeo dual digital/análogico Mini-DV/S-VHS JVC SR-VS10U, o material foi transferido, através de interface digital IEEE 1394 (*Firewire)*, para uma ilha de edição não-linear iFinish V60 versão 3.2, Media 100. O material foi segmentado e rotulado manualmente nas fronteiras de produção dos itens pronunciados tomando-se como referência o tom de 1kHz, resultando em um conjunto 165 fragmentos audiovisuais. Cada fragmento passou por processo de separação das trilhas de áudio e vídeo, resultando em dois produtos: um arquivo de áudio PCM (amostrado a 48 kHz, 16 bits/amostra e codificação linear sem compressão), e um conjunto de imagens referentes aos quadros do fragmento de vídeo correspondente. Segundo as características do padrão NTSC, cada segundo de gravação corresponde à captura de aproximadamente 30 quadros, que foram digitalizados em imagens com resolução 720 x 486 pixels. As imagens foram armazenadas em formato Microsoft Windows BMP sem compressão. A extração de quadros dos fragmentos de vídeo obtidos totalizaram aproximadamente 13.000 imagens.

3.4 Análise do Áudio

Os arquivos de áudio resultantes do processo de segmentação e rotulação do corpus audiovisual passaram por processo de análise e transcrição fonética temporizada, que permitiu associar as imagens capturadas aos fones pronunciados pela apresentadora.

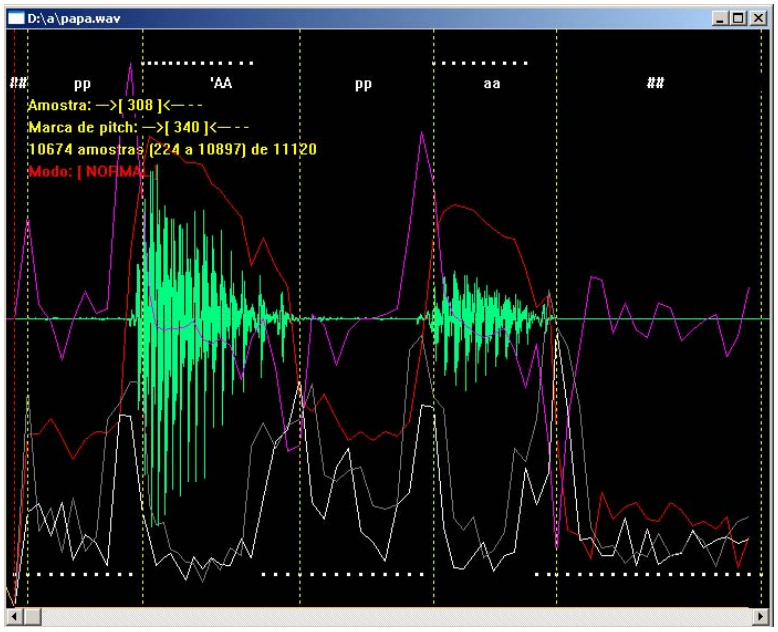


Fig. 3.3: Inspeção visual do áudio a partir de sua forma de onda

O processo de transcrição fonética foi realizado com o auxílio de ferramentas de apoio utilizadas no desenvolvimento do conversor texto-fala "CPqD Texto Fala", produto comercial da Fundação CPqD. Em particular, duas ferramentas de análise de áudio foram utilizadas: uma ferramenta de segmentação automática que identifica as fronteiras de segmentos sonoros da fala e um aplicativo de análise visual do sinal de áudio. As ferramentas utilizadas não são soluções comerciais e são utilizadas para suporte ao processo de desenvolvimento do "CPqD Texto Fala". O processo completo de transcrição contou com o apoio técnico da equipe responsável por este desenvolvimento.

Inicialmente os arquivos de áudio PCM originalmente capturados a 48kHz foram sub-amostrados para a taxa de 16kHz, resultando em arquivos PCM com 16 bits/amostra e codificação linear sem compressão.

Em seguida, cada arquivo de áudio foi submetido a um processo de segmentação automática que identifica as fronteiras de início e fim da produção acústica de cada segmento sonoro presente na locução analisada, a partir da transcrição textual do conteúdo pronunciado. No entanto, a taxa de acertos obtida durante o processo segmentação automática não foi satisfatória, mostrando-se necessária a revisão auditiva e visual, através das formas de onda do sinal de aúdio de cada item considerado. Tal resultado é atribuído ao fato de que a ferramenta de segmentação automática é treinada para uma voz feminina específica, diferente da presente nos arquivos analisados.

Prosseguiu-se assim com a análise visual da forma de onda do sinal de áudio através de ferramenta que permite posicionar interativamente as marcas de segmentação. A ferramenta utilizada também disponibiliza a visualização de curvas informativas sobre a variação de energia do sinal de áudio correspondente e sobre o comportamento de parâmetros característicos do sinal de áudio. A Figura 3.3 mostra a tela da ferramenta utilizada para inspeção visual do sinal de áudio, mostrando a forma de onda para o logatoma ['papɐ]. As linhas tracejadas verticais em cor clara representam as marcas de segmentação para este exemplo.

Para cada arquivo de áudio analisado, o processo de transcrição fonética resultou na geração de um arquivo texto associado contendo a sequência de fones presentes no item pronunciado, acompanhados de seus instantes de início.

3.5 Análise do Vídeo

Conforme descrito na Seção 3.3, um dos produtos da segmentação do corpus audiovisual foi a obtenção de aproximadamente 13.000 imagens correspondentes a quadros de vídeo de cada um dos 165 fragmentos audiovisuais capturados. No entanto, como discutido na Seção 3.2.1, apenas a parcela das imagens provenientes da captura em vídeo da locução de logatomas foi processada visando a extração de visemas para a construção da base de imagens. A partir da transcrição fonética temporizada dos arquivos de áudio correspondentes, foi possível associar cada uma das imagens extraídas do corpus à produção acústica de um determinado fone, ou à ausência de locução (silêncio).

Nas seções seguintes são descritos os processos de seleção dos 34 visemas que constituem a base de imagens e os procedimentos de processamento e extração de características visuais de suas imagens correspondentes.

3.5.1 Seleção das Imagens

A seleção dos visemas da base de imagens foi realizada com o auxílio de uma ferramenta desenvolvida em C++ para visualização das imagens extraídas do corpus audiovisual levando-se em consideração a transcrição fonética resultante do processo de análise do aúdio (Seção 3.4). Esta ferramenta foi especialmente desenvolvida no contexto deste trabalho.

A ferramenta permite a análise visual quadro a quadro de um determinado fragmento de vídeo, associando os quadros à produção acústica de um determinado fone. A título de exemplo, a Figura 3.4 ilustra como estas informações são disponibilizadas pela ferramenta. A figura mostra um subconjunto de 10 quadros extraídos do fragmento de vídeo capturado durante a pronúncia do logatoma ['papɐ], dando ênfase aos quadros capturados durante a produção acústica do segundo fone [p]. A linha temporal mostra os instantes em que as imagens foram capturadas a partir do início do fragmento de vídeo, obedecendo a frequência de captura de aproximadamente 30 quadros por segundo. A partir da transcrição fonética temporizada resultante da análise do áudio correspondente, é possível rotular cada uma das imagens com o fone que elas representam.

O objetivo da seleção de imagens visando a construção de uma base de visemas é eleger a imagem que melhor representa a postura labial característica associada a um visema dependente de contexto.

Neste trabalho, considerou-se que a postura labial característica é realizada no instante em que ocorre uma parada, com eventual mudança de direção, na excursão ou trajetória definida pelos lábios. Visualmente, uma vez realizada a postura labial característica, a trajetória traçada pelos lábios é alterada, movendo-se para realizar a postura caracterísitca do próximo segmento.

A Figura 3.4 ilustra este processo, exemplificando a escolha da imagem representante para o visema $< p_1 >$ da Tabela 3.3. Inicialmente, seleciona-se do corpus um do itens pronunciados que contenha um dos contextos fonéticos associados a este visema, como por exemplo o logatoma ['papɐ], que contém o contexto fonético [apɐ]. Em seguida, analisa-se visualmente todos os quadros correspondentes ao fone pertencente ao grupo homofema associado ao visema em questão, no caso, o fone

[p]. Elege-se como imagem representante do visema aquela que visualmente indica a expressão máxima de articulação dentre os quadros associados à produção deste fone. Na figura, dentre os quatro quadros associados ao fone [p], o quadro selecionado está destacado por um retângulo que o circunda. No caso do fone [p], a imagem que melhor representa o visema dependente de contexto $< p1 >$ foi associada ao ponto máximo de contratura dos lábios antes da explosão sonora característica deste tipo de segmento acústico.

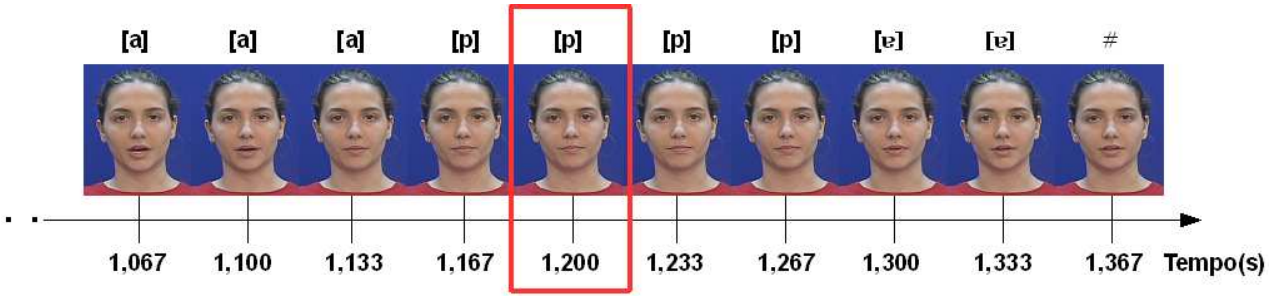


Fig. 3.4: Seleção de um visema

Os 22 visemas consonantais da Tabela 3.3, bem como os visemas $< i_1 >$, $< i_2 >$, $< \text{ɪ} >$, $< a >$, $< \text{ɐ} >$, $< u >$, $< \text{ʊ} >$ da Tabela 3.4, foram extraídos dos logatomas do tipo $'CV_1CV_2$ definidos na Seção 3.2.1. Já os visemas $< e >$, $< \text{ɛ} >$, $< o >$, $< \text{ɔ} >$ da Tabela 3.4, foram extraídos dos fragmentos de vídeo provenientes da locução dos ditongos do tipo V_1V_2, também definidos na Seção 3.2.1. Finalmente, o visema representante da posição de repouso (silêncio) foi extraído de um trecho de vídeo sem locução.

3.5.2 Registro das Imagens

As imagens extraídas do corpus passaram por um processo de alinhamento em relação a uma imagem de referência. O objetivo dessa operação é corrigir variações de enquadramento da face no vídeo, fazendo com que os elementos da face tais como olhos, nariz e lábios estejam registrados uns em relação aos outros, apresentando orientação, escala e posicionamento casados.

Uma das premissas assumidas para o processo de alinhamento das imagens foi que as imagens não apresentavam alterações significativas da pose da cabeça, hipótese garantida pelo monitoramento do enquadramento da face durante o processo de captura do corpus.

O problema da uniformização do enquadramento e posição dos elementos da face entre as diversas imagens extraídas do corpus pode ser tratado como um problema de registro, em que se busca o alinhamento geométrico entre duas imagens de uma mesma cena, capturadas em instantes diferentes.

A abordagem adotada elegeu, inicialmente, uma das imagens extraídas do corpus como imagem de referência. Tal imagem foi escolhida por possuir características como: nitidez, posição da cabeça centralizada na imagem e ausência de rotação da cabeça no plano da imagem. A imagem de referência assume posteriormente, durante o processo de síntese, o papel de face-base, conceito definido no Capítulo 4.

O próximo passo consistiu na detecção manual dos pontos característicos mostrados na Figura 3.5 na imagem de referência e nas 34 imagens selecionadas para formar a base de imagens. Os pontos característicos correspondem aos cantos internos dos olhos (OD e OE), pontos de união dos

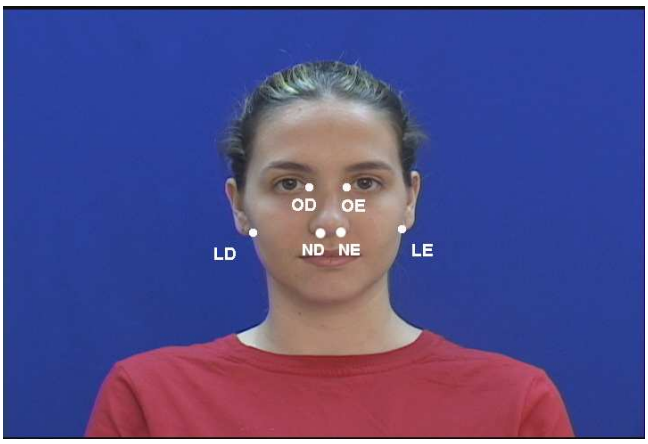


Fig. 3.5: Pontos característicos para registro

lóbulos das orelhas com a face (LD e LE) e narinas (ND e NE). Estes pontos foram escolhidos por serem facilmente identificáveis através de inspeção visual e minimamente afetados pelo processo de produção da fala, sendo bons indicadores de movimentações rígidas da face na imagem.

O procedimento de detecção visual dos pontos e a aplicação do algoritmo de registro foram realizados com o auxílio de uma ferramenta desenvolvida em C++, caracterizada como uma ferramenta de auxílio ao processamento das imagens da base. Esta ferramenta, desenvolvida especialmente no contexto deste trabalho, permite que um usuário realize manualmente a marcação dos pontos utilizados como referência (Figura 3.5) para o processo de registro. Para este procedimento, o aplicativo disponibiliza recursos de *zoom* na imagem, facilitando o processo de detecção visual e possibilitando maior precisão na detecção dos pontos característicos. Após a detecção dos mesmos, todos os cálculos e transformações necessárias nas imagens a serem registradas são realizadas automaticamente.

Após a detecção e medição das coordenadas dos pontos característicos das imagens analisadas, o processo de registro foi implementado através de uma transformação espacial ou *warping*. O *warping* de uma imagem pode ser definido como uma operação de distorção desta imagem, através da redefinição da relação espacial entre seus pontos. Desta maneira, o processo de registro consiste na distorção de uma imagem de maneira que seus pontos característicos originais sejam reposicionados nas coordenadas dos pontos característicos correspondentes da imagem de referência.

Considerando-se o pequeno número de pontos característicos considerados neste trabalho, a função de *warping* consiste em uma função capaz de interpolar dados esparsos no espaço. Neste tipo de problema, os pontos característicos são denominados pontos de controle da função de interpolação, ou pontos-âncora do processo de transformação geométrica espacial.

Vários são os possíveis métodos de registro de imagens e transformações geométricas que podem ser utilizadas com este propósito (ZITOVÁ; FLUSSER, 2003). Em particular, neste trabalho buscou-se identificar a abordagem de *warping* que fosse mais apropriada a imagens faciais, tendo-se como base a literatura existente relacionada a animação facial e manipulação de expressões faciais. Assim, a partir de trabalhos como (ARAD et al., 1994), (PIGHIN et al., 1998) e (EDGE; MADDOCK, 2003), optou-se por utilizar funções de base radial (RBF - *Radial Basis Functions*). Esta categoria de funções é popularmente aplicada em problemas de interpolação de dados esparsos, e são comumente aplicadas no *warping* de imagens de superfícies irregulares (AMIDROR, 2002), (RUPRECHT; MÜLLER, 1995),

(ZITOVÁ; FLUSSER, 2003).

Funções de base radial são funções de interpolação globais, ou seja, todos os pontos-âncora a serem interpolados são considerados durante a determinação dos parâmetros da função de *warping*. Em outras palavras, o valor da função de transformação para cada *pixel* da imagem de entrada é função de todos os pontos-âncora. Em particular, a expressão "radial" reflete a propriedade de que este valor é função apenas da distância do ponto considerado aos pontos-âncora, não importando a localização do mesmo na imagem.

A seguir descreve-se a formulação empregada para o *warping* de imagens por meio de funções de base radial. Além do registro de imagens da base de imagens, esta formulação foi também empregada durante o processo de metamorfose entre imagens durante a síntese dos quadros da animação, processo descrito no Capítulo 4.

Transformação Espacial de Imagens Utilizando Funções de Base Radial

A função de *warping* é implementada de maneira que o conjunto de pontos-âncora da imagem inicial seja obrigatoriamente mapeado para o conjunto de pontos-âncora correspondente da imagem final. Tal associação é ilustrada pelo mapa de correspondência mostrado na Figura 3.6. Nesta figura, adota-se a seguinte notação:

- uma imagem digital é considerada o resultado de um processo de amostragem e quantização de uma cena real $f(x, y)$, representada por uma matriz de M linhas e N colunas, em que coordenadas (x, y) são quantidades discretas e inteiras, e $f(x, y)$ fornece o valor do *pixel* correspondente;
- a imagem $f(x, y)$ é a imagem inicial a ser transformada e todos os valores de *pixel* em seu domínio são conhecidos;
- a imagem $g(x', y')$, com coordenadas (x', y'), é a imagem final transformada, cujos valores são determinados pela aplicação da função de *warping* à imagem inicial;
- o processo de *warping* é guiado pela associação de k pontos-âncora $P_i = (x_i, y_i)$ da imagem inicial a k pontos-âncora $Q_i = (x'_i, y'_i)$ da imagem resultante final, onde $i = 1...k$.

A transformação de *warping* implementada pode então ser definida como um problema de entrada e saída, onde:

- **Entrada:**
  - imagem $f(x, y)$, com dimensões *M*x*N*, com *pixels* de coordenadas (x, y);
  - k pares de pontos (P_i, Q_i), onde P_i e $Q_i \in \Re^2, i = 1...k$.
- **Saída:** imagem $g(x', y')$, com dimensões *M*x*N* e *pixels* de coordenadas (x', y'), onde ($i = 1...k$):

$$\begin{cases} x' = r(x, y) \\ y' = s(x, y) \\ x'_i = x_i \\ y'_i = y_i \end{cases} \tag{3.1}$$

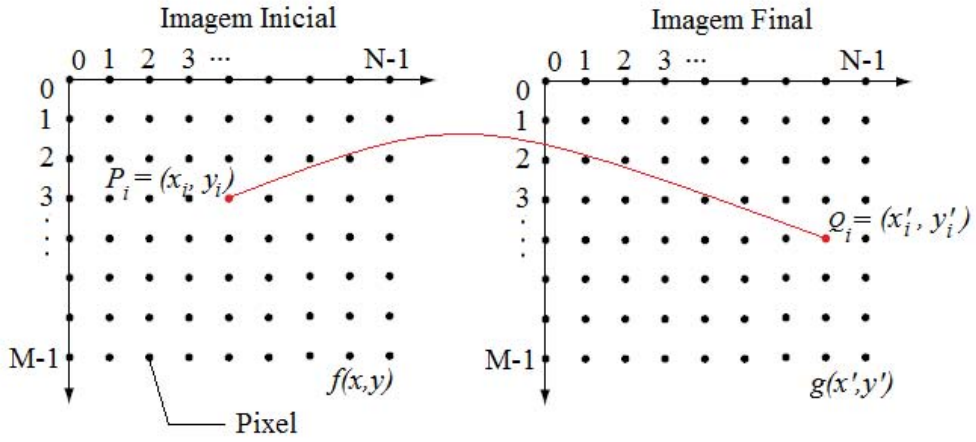


Fig. 3.6: Mapa de correspondência da imagem inicial à imagem final a ser transformada

A função de interpolação adotada é uma combinação linear de funções radiais, que possuem como principal característica serem funções da distância $d(P, P_i)$ do *pixel* genérico $P = (x, y)$ ao ponto-âncora P_i:

$$\begin{cases} r(x,y) = r(P) = b_m(P) + \sum_{i=1}^{k} \alpha_i \phi_i(d(P,P_i)) \\ s(x,y) = s(P) = b_m(P) + \sum_{i=1}^{k} \beta_i \phi_i(d(P,P_i)) \end{cases} \tag{3.2}$$

onde:

- $b_m(P)$ é um polinômio de grau m;
- $d(P, P_i)$ é a distância entre o ponto genérico P e o i-ésimo ponto âncora P_i:

$$d(P,P_i) = \sqrt{(x-x_i)^2 + (y-y_i)^2} \tag{3.3}$$

- $\phi_i(d(P, P_i))$ é uma função radial que tem como característica afetar igualmente todos os pontos equidistantes de P_i;
- α_i e β_i são coeficientes escalares.

O polinômio $b_m(P)$ adotado é um polinômio linear ($m = 1$) que possui como função permitir que a função de interpolação seja capaz de mapear transformações lineares (não modeladas por somatórias puras de funções radiais (ARAD et al., 1994)), possuindo a seguinte formulação:

$$b_1(P) = a_1 + a_2 \cdot x + a_3 \cdot y \tag{3.4}$$

Dentre as diversas formulações possíveis para funções radiais, o sistema adota uma função multi-quadrática, apontada em (RUPRECHT; MÜLLER, 1995) como efetiva e computacionalmente eficiente:

$$\phi_i(d(P,P_i)) = (d(P,P_i)^2 + r_i^2)^{\mu}, r_i > 0, \mu \neq 0 \tag{3.5}$$

A partir da análise realizada em (RUPRECHT; MÜLLER, 1995), μ foi definido como 0.5 e um valor diferente para r_i foi usado para cada P_i, sendo este definido como a distância ao ponto-âncora vizinho mais próximo:

$$\begin{cases} r_i = min_{i \neq j}(d(P_i, P_j)) \\ i = 1...k \\ j = 1...k \end{cases} \tag{3.6}$$

Os coeficientes a_1, a_2, a_3, α_i e β_i, $i = 1...k$, são determinados através da solução do sistema de equações lineares resultantes das condições de contorno impostas pelo mapa de correspondência e pela condição de precisão polinomial para $m = 1$ (que atribui à função de interpolação a capacidade de reproduzir polinômios lineares).

Para $x' = r(x, y)$, tais condições são:

$$\begin{cases} x'_i = x_i \\ \sum_{i=1}^{k} \alpha_i = 0 \\ \sum_{i=1}^{k} \alpha_i \cdot x_i = 0 \\ \sum_{i=1}^{k} \alpha_i \cdot y_i = 0 \end{cases} \tag{3.7}$$

Para $y' = s(x, y)$, tais condições são:

$$\begin{cases} y'_i = y_i \\ \sum_{i=1}^{k} \beta_i = 0 \\ \sum_{i=1}^{k} \beta_i \cdot x_i = 0 \\ \sum_{i=1}^{k} \beta_i \cdot y_i = 0 \end{cases} \tag{3.8}$$

Cada transformação espacial entre duas imagens envolve a solução de um sistema de equações lineares com $k + 3$ equações e $k + 3$ incógnitas. Os sistemas de equações resultantes das condições representadas pelas equções 3.7 e 3.8 são apresentados no Apêndice B, através dos sistemas B.1 e B.2, respectivamente.

Mapeamento Inverso de *Pixels* em Transformações Espaciais

Em transformações espaciais, é possível calcular a correspondência dos pixels da imagem de entrada para pixels da imagem de saída, ou vice-versa, caracterizando duas possíveis estratégias: mapeamento direto ou mapeamento inverso (JÄHNE, 2005).

Através do mapeamento *direto*, um *pixel* da imagem de entrada é mapeado na imagem de saída. Nesta técnica, *pixels* da imagem de entrada podem ser mapeados fora do domínio desejado da imagem de saída e/ou mapeados para regiões entre *pixels* da imagem de saída. A aplicação desta técnica pode acarretar "buracos" na imagem de saída, resultante de *pixels* nunca mapeados, ou um valor pode ser atribuído mais de uma vez a um mesmo *pixel* de saída.

Neste trabalho, as transformações espaciais foram implementadas aplicando-se a técnica de mapeamento *inverso* para obtenção da imagem final. Esta técnica percorre o domínio de pixels da imagem de saída aplicando a transformada inversa e determinando-se o pixel correspondente na imagem de entrada. O valor do pixel da entrada é então copiado para a imagem de saída. Esta técnica garante que todos os pixels de saída são computados, já que todos os *pixels* da imagem de saída são percorridos sequencialmente. Nesta abordagem, os *pixels* da imagem de saída mapeados para posições intermediárias entre *pixels* da imagem de entrada são geralmente calculados a partir da interpolação dos valores destes *pixels*. Neste trabalho, foi adotada a técnica de interpolação bilinear. Adicionalmente, foi adotada a estratégia em que *pixels* mapeados para fora do domínio da imagem de entrada são simplesmente preenchidos com um valor *pixel* pré-definido (a implementação adotada neste trabalho utiliza o valor "0"). Apenas um pequeno conjunto *pixels* localizados nas bordas das imagens processadas e fora da região de interesse(vide Seção 3.5.3) se encaixam nesta condição.

3.5.3 Região de Interesse e Pontos-âncora do Visema

Após o alinhamento das 34 imagens extraídas do corpus em relação a uma imagem de referência, o próximo passo foi a extração de uma região de interesse que compreende a região dos lábios e queixo das imagens faciais consideradas. Esta operação foi realizada com o auxílio da mesma ferramenta utilizada para o processo de registro das imagens (Seção 3.5.2). O aplicativo desenvolvido realiza automaticamente o processo de extração da região de interesse na face, e oferece recursos para detecção manual dos pontos-âncora utilizados como guias para o processo de síntese da animação. Uma vez que nesta etapa do processo as imagens estão alinhadas, a extração da região de interesse consiste em uma simples operação de recorte de uma mesma região retangular em todas as imagens da base. A Figura 3.7 apresenta a região considerada para formação da base de imagens.

A extração desta região reflete a estratégia de síntese adotada, que funde os visemas a uma face-base (Capítulo 4). Esta estratégia permite o armazenamento na base de imagens com dimensões de 200 x 150 pixels, reduzindo em cerca de 10 vezes o tamanho em bytes da base de imagens em relação ao armazenamento das imagens completas.

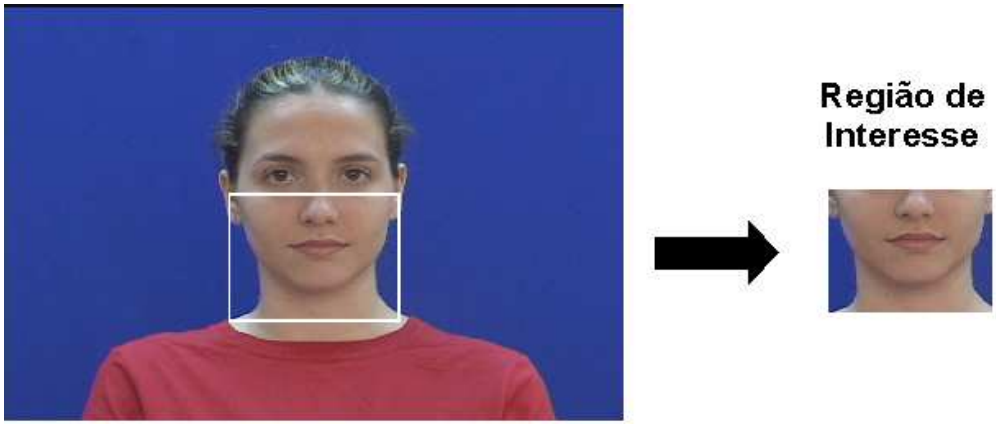


Fig. 3.7: Extração da Região de Interesse

Após o processo de extração da região de interesse, foi realizada a medição dos pontos-âncora utilizados pelo processo de síntese da animação. O termo "ponto-âncora", como discutido na seção

anterior, reflete a idéia de que tais pontos devem permanecer estáticos durante as sucessivas transformações geométricas realizadas no processo de metamorfose entre visemas. Mais uma vez, o processo de medição dos pontos-âncora e o armazenamento de suas coordenadas foram realizados de maneira semi-automática pela ferramenta de auxílio ao processamento das imagens da base de imagens.

Os pontos-âncora considerados neste trabalho são mostrados na Figura 3.8.

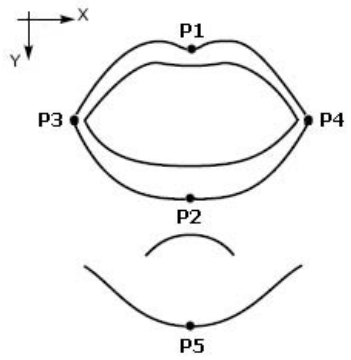


Fig. 3.8: Pontos-âncora

3.6 Resultados

O resultado final do processo de construção da base de imagens pode ser visualizado através da Figura 3.9, que contempla 1 visema de repouso (silêncio), 22 visemas consonantais (Tabela 3.3) e 11 visemas vocálicos (Tabela 3.4).

Além do conjunto de imagens mostrado na Figura 3.9, a base de imagens inclui uma série de informações de rótulo associadas a cada uma das imagens da base. As propriedades associadas a cada imagem são:

- caminho para o arquivo da imagem;
- identificador do fragmento de vídeo do qual a imagem foi extraída;
- identificador do visema de acordo com as tabelas 3.3 e 3.4;
- pentafone (sequência de 5 fones) que indica o contexto fonético no qual a imagem foi capturada incluindo: 2 fones adjacentes à esquerda do fone central, fone central, 2 fones adjacentes à direita do fone central;
- coordenadas x e y dos 5 pontos-âncora medidos para o visema (Figura 3.8).

3.7 Comentários Finais

O presente capítulo apresentou o processo de modelagem da face adotado neste trabalho, descrevendo a metodologia empregada para captura do corpus audiovisual e as etapas de processamento das imagens extraídas do corpus visando a construção de uma base de imagens.

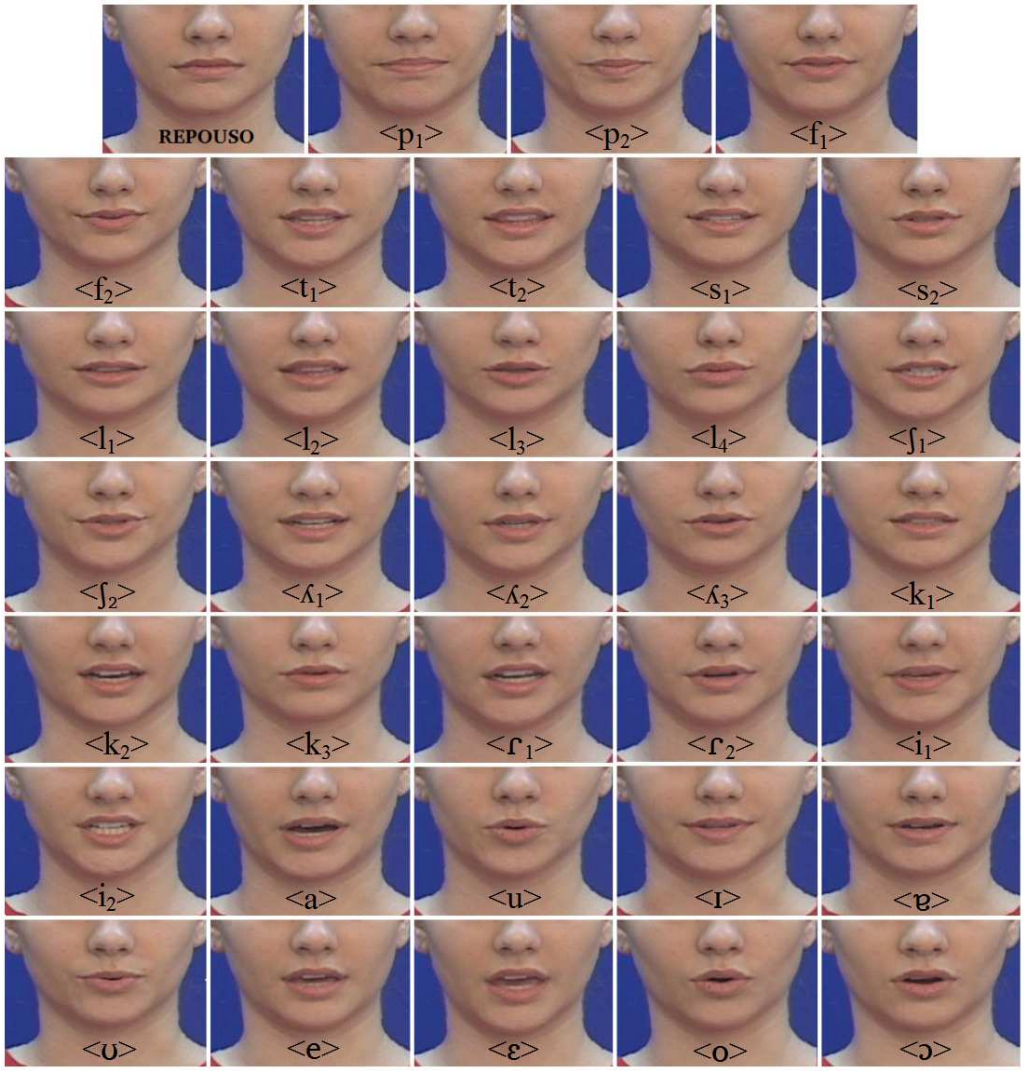


Fig. 3.9: Base de Imagens com 34 visemas

Este trabalho tem como foco principal a reprodução vídeo-realista dos movimentos articulatórios visíveis da fala e, portanto, a base de imagens implementada não inclui quaisquer dados ou elementos visuais adicionais que permitam a reprodução de diferentes poses da cabeça, movimentos de comunicação não-verbal ou a reprodução de expressões faciais relacionadas a emoções.

Os princípios utilizados para construção da base de imagens refletem a modelagem da articulação visível adotada neste trabalho, que contempla os efeitos da coarticulação, através da adoção de visemas dependentes de contexto para o Português do Brasil.

A base de imagens resultante é composta de 34 visemas dependentes de contexto, caracterizados por imagens que compreendem a região dos lábios e queixo, descartando as informações visuais do restante da face. Esta característica reflete a estratégia de síntese baseada em visemas fundidos a uma face-base. Esta abordagem reduz drasticamente o custo de armazenagem em memória das imagens da base e, como será apresentado nos capítulos que se seguem, tornam o processo de síntese mais rápido.

Tal abordagem é especialmente vantajosa para dispositivos com capacidade limitada de memória, permitindo que a base de imagens resida na memória de dispositivos tais como *smartphones*, decodificadores de TV digital e assistentes pessoais (PDAs - *Personal Digital Assistants*). Outra vantagem é que a base possa ser transmitida rapidamente através de canais com baixas taxas de transmissão.

Outra característica a ser ressaltada é a relativa simplicidade do processo de seleção de imagens do corpus e dos algoritmos de processamento digital de imagens utilizados para formatação das imagens da base, quando comparados aos sistemas apresentados no Capítulo 2. Esta característica permite conceber aplicações em que o usuário criaria facilmente uma base de imagens de sua própria face de maneira semi-automática, como apresentada neste trabalho, ou totalmente automática a partir da implementação de algoritmos de detecção automática de elementos faciais.

Capítulo 4

Síntese da Animação

4.1 Introdução

Este capítulo descreve a abordagem de síntese da animação facial 2D adotada neste trabalho.

A Figura 4.1 apresenta o processo de síntese que possui, como parâmetro de entrada, a transcrição fonética do áudio correspondente à fala a ser visualmente animada. Como discutido no Capítulo 2 (Seção 2.3.1), a transcrição fonética temporizada é composta pela sequência de fones que compõem a locução e suas respectivas durações, e pode ser obtida através de processos como: análise manual do áudio, aplicação de algoritmos de segmentação automática da voz ou pode ser o resultado intermediário de um processo de conversão texto-fala.

Inicialmente, o capítulo descreve as informações essenciais que obrigatoriamente devem constar da transcrição fonética temporizada e que possibilitam a síntese da animação facial em sincronia e harmonia com o áudio correspondente (Seção 4.2). Também nesta seção, descreve-se como a informação temporal proveniente da transcrição é interpretada, visando a determinação dos instantes associados às poses-chave da animação, caracterizadas por visemas da base de imagens.

Além do processamento da informação temporal, a transcrição fonética temporizada é processada e analisada resultando na determinação de uma sequência de visemas dependentes de contexto cor-

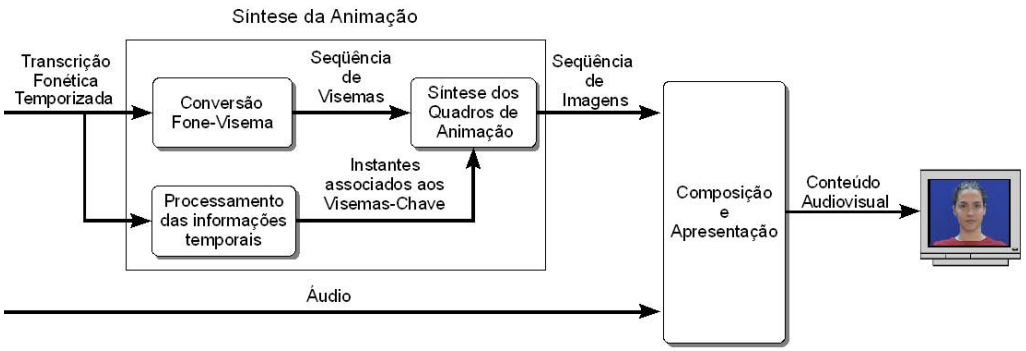


Fig. 4.1: Entrada e Saída do Processo de Síntese da Animação Facial 2D

respondente. Conforme apresentado na Seção 4.3, esta etapa compreende importantes aspectos da modelagem articulatória adotada. Como resultado fundamental desta etapa, obtém-se a sequência de poses-chave da animação, ou simplesmente, visemas-chave.

Após a definição da sequência de visemas-chave e seus instantes de realização, prossegue-se com a síntese dos quadros da animação. A descrição deste processo é apresentada na Seção 4.4. A síntese dos quadros de animação é baseada em três passos principais: a determinação de uma curva de interpolação não-linear que guia a variável temporal do processo de metamorfose, a utilização de funções de base radial (RBF - *Radial Basis Functions*) para a implementação das deformações espaciais dos visemas-chave e a fusão dos visemas sintetizados a uma face base. A combinação das estratégias adotadas para cada um destes passos, tornam única a solução apresentada neste trabalho.

4.2 Transcrição Fonética Temporizada

A transcrição fonética de um determinado áudio ou texto consiste na representação por meio de símbolos fonéticos da sequência de segmentos da fala associada. Visando a animação visual da fala, a transcrição fonética do aúdio correspondente deve estar acompanhada de informações temporais que possibilitem determinar as fronteiras de cada fone e a duração dos mesmos.

A Figura 4.2 exemplifica como a transcrição fonética temporizada fornece informações ao processo de síntese da animação. Na figura, uma sequência de n fones F_i, $i = 1, 2...n$ tem seus instantes de início determinados pelos tempos correspondentes t_i. O intervalo de produção acústica de um determinado fone F_i é determinado pela diferença $(t_{i+1} - t_i)$. A duração da animação deve coincidir com duração da locução dada por $(t_{n+1} - t_0)$, representada na figura por ΔT.

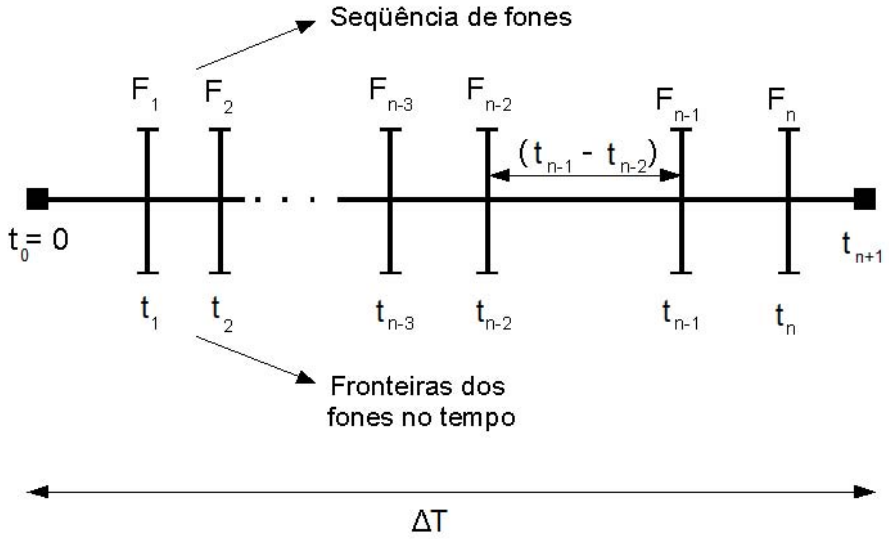


Fig. 4.2: Transcrição fonética temporizada

A partir das informações temporais fornecidas pela transcrição fonética temporizada, é possível determinar o número de quadros da animação final e os instantes aos quais devem estar associados os visemas-chave da base de imagens.

A animação correspondente à locução de uma sequência de segmentos, é realizada associando-se visemas-chave aos instantes de ocorrência dos alvos articulatórios dos fones da sequência. Um alvo articulatório é definido como a postura articulatória que caracteriza a conformação típica do trato vocal associada à locução de um determinado segmento. No contexto deste trabalho, os visemas-chave representam alvos articulatórios e são associados aos instantes em que ocorre uma parada, com eventual mudança de direção, na excursão ou trajetória definida pelos lábios. Dessa maneira, o processo de síntese consiste na geração de quadros intermediários entre visemas-chave, visando reproduzir a trajetória traçada pelos lábios ao se moverem de uma postura característica para outra, correspondente ao próximo segmento da locução.

Neste trabalho considera-se que o alvo articulatório de um determinado fone ocorre exatamente no meio do intervalo de duração do fone, como mostrado pela Figura 4.3. Assim, considerando-se dois fones adjacentes F_i e F_{i+1}, o alvo articulatório A_i do fone F_i é associado ao instante:

$$A_i = t_i + \frac{t_{i+1} - t_i}{2} = \frac{t_i + t_{i+1}}{2} \tag{4.1}$$

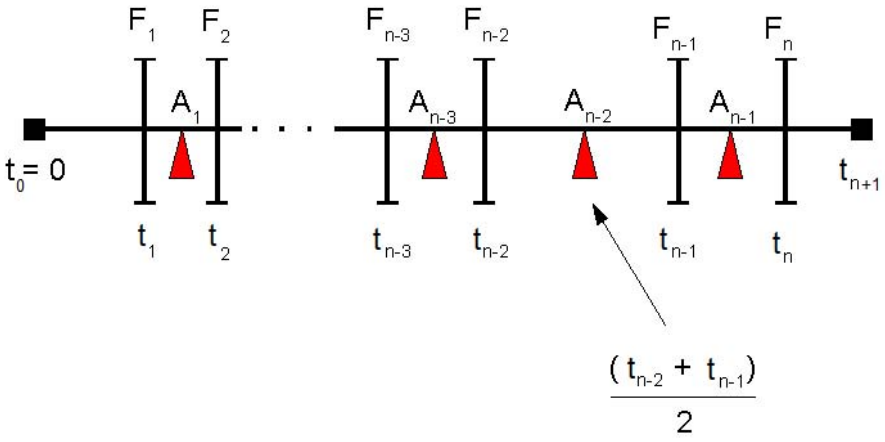


Fig. 4.3: Instantes associados aos alvos articulatórios

4.3 Conversão Fone-Visema

Além da determinação dos instantes correspondentes aos alvos articulatórios (Seção 4.2), a transcrição fonética temporizada é processada visando a obtenção de uma sequência de visemas dependentes de contexto correspondente às poses-chave da animação. Como discutido no Capítulo 3 (Seção 3.2), os visemas considerados neste trabalho são caracterizados por serem visemas dependentes de contexto e são apresentados nas tabelas 3.3 e 3.4.

O processo de conversão fone-visema (Figura 4.1) pode ser resumido através dos passos descritos a seguir.

- **PASSO 1 - Identificação dos grupos homofemas:** a sequência de fones é percorrida e cada fone é substituído pelo fone representante de seu grupo homofema segundo as tabelas 4.1 e 4.2.

Tab. 4.1: Homofemas consonantais e fones representantes.

Grupo Homofema	Fone Representante
[p,b,m]	[p]
[f,v]	[f]
[t,d,n]	[t]
[s,z]	[s]
[l]	[l]
[k,g]	[k]
[ʃ, ʒ]	[ʃ]
[ʎ, ɲ]	[ʎ]
[ɾ],[ɣ]	[ɾ]

Tab. 4.2: Homofemas vocálicos e fones representantes.

Grupo Homofema	Fone Representante
[i,ĩ]	[i]
[e,ẽ]	[e]
[ɛ]	[ɛ]
[a,ɐ̃]	[a]
[ɔ]	[ɔ]
[o,õ]	[o]
[u,ũ]	[u]
[ɪ]	[ɪ]
[ɐ]	[ɐ]
[ʊ]	[ʊ]

- **PASSO 2 - Conversão dos fones vocálicos:** conforme mostrado na tabela 3.4, a maioria dos visemas vocálicos considerados são independentes do contexto fonético em que eles são produzidos. Assim, a sequência de fones resultantes do **PASSO 1** é percorrida, e os fones vocálicos {[i, ĩ, e, ẽ, ɛ, a, ɐ̃, ɔ, o, õ, u, ũ, ɪ, ɐ, ʊ]} são imediatamente associados aos visemas correspondentes da tabela 3.4. No entanto, na presença de um fone vocálico [i], o trifone que tem como fone central o fone [i] deve ser considerado. Na presença dos contextos [tit] e [ʃiʃ], o visema associado é <i_2>. Para todos os outros contextos, o visema associado é <i_1>.

- **PASSO 3 - Simplificação da cadeia de fones através de substituição de fones vocálicos:** após a conversão dos fones vocálicos, a cadeia de fones resultante do **PASSO 1** é mais uma vez percorrida, e os fones vocálicos são novamente substituídos. Tal operação simplifica a cadeia de fones a ser processada no passo seguinte, que consiste na conversão dos fones consonantais.

Tab. 4.3: Tabela de substituição de fones vocálicos.

Fone Encontrado	Fone de Substituição
[i, ĩ, ɪ]	[i]
[a, ɛ, ɔ, ɐ̃, ɐ, e, ẽ]	[a]
[u, o, õ, ũ, ʊ]	[u]

- **PASSO 4 - Análise dos fones consonantais:** a conversão de fones consonantais em visemas exige a análise do contexto fonético em que eles foram produzidos. Em particular, para fones consonantais, os visemas dependentes de contexto adotados neste trabalho são obtidos a partir do mapeamento dos diversos contextos fonéticos possíveis para o Português do Brasil em contextos do tipo $\#CV$ e V_1CV_2 da coluna "Contextos Fonéticos" da Tabela 3.3, onde (vide Seção 3.2):

  - C ∈{[p,f,t,s,l,k,ʃ,ʎ,l,(ɣ)ɾ]};
  - V,V_1 ∈{[i,a,u]};
  - V_2 ∈{[ɪ,ɐ,ʊ]}.

  Considerando-se a simplificação de fones vocálicos realizada no **PASSO 3**, obtém-se:

  V,V_1,V_2 ∈{[i,a,u]}.

  A Seção 4.3.1 define as diretrizes assumidas para o mapeamento adotado, apresentado na Tabela 4.9.

- **PASSO 5 - Conversão dos fones consonantais:** após o mapeamento de contextos fonéticos realizada no **PASSO 4**, é possível realizar a conversão de fones consonantais nos seus visemas correspondentes. Tal conversão é realizada, primeiramente, considerando-se a simplificação de fones vocálicos aplicada à coluna "Contextos Fonéticos" da Tabela 3.3 e, em seguida, fazendo-se a associação direta entre os novos contextos obtidos e os contextos mapeados para cada fone

consonantal da transcrição fonética temporizada. A Sessão 4.3.1 descreve as diretrizes e os conceitos fundamentais utilizados neste trabalho para a adoção do mapeamento proposto pela Tabela 4.9.

4.3.1 Mapeamento de Contextos Fonéticos para Fones Consonantais

A conversão de fones consonantais em visemas exige o mapeamento dos diversos contextos fonéticos possíveis de ocorrência destes fones para o Português do Brasil.

Para a determinação destes contextos é necessário, primeiramente, considerar as particularidades desta língua relacionadas aos padrões silábicos que envolvem segmentos, ou fones consonantais.

Na produção da fala, as sílabas são o resultado da agregação de fones para formação de unidades mínimas de pronúncia, que podem ser pronunciadas em uma só emissão de voz. Assim, a palavra "pata" pode ser desmembrada nas sílabas /pa/ e /ta/, não fazendo sentido o desmembramento /p/, /a/, /t/ e /a/, já que os fonemas /p/ e /t/ são impronunciáveis isoladamente.

Toda locução é constituída por pelo menos uma sílaba. A estrutura da sílaba pode ser considerada como tendo três partes: ataque, núcleo e coda. Das três partes, apenas o núcleo é obrigatório. Já o ataque e a coda são opcionais. No Português do Brasil, o ataque e a coda podem ser formados por até dois fones consonantais cada um, enquanto o núcleo pode ser formado por até três fones vocálicos. Com base nestas regras, as tabelas 4.4, 4.5, 4.6 e 4.7 apresentam exemplos dos possíveis padrões silábicos para o Português do Brasil. Nestas tabelas, a primeira coluna apresenta o padrão silábico exemplificado, a segunda coluna apresenta um exemplo de palavra em que ocorre tal padrão e na terceira coluna são apresentados os fonemas consonantais considerados (como por exemplo /p/). A representação textual dos fonemas é realizada utilizando-se o Alfabeto Fonético Internacional (INTERNATIONAL PHONETIC ASSOCIATION, 1999). Uma análise mais aprofundada e abrangente de tais padrões silábicos é apresentada em (DE MARTINO, 2005).

Adicionalmente, a Tabela 4.8 apresenta exemplos de alguns possíveis encontros consonantais que, neste trabalho, são considerados pseudo-encontros consonantais pois, para a animação visual da fala, será assumida a existência da vogal epentética /ɪ/ entre os segmentos consonantais. Assim, por exemplo, a palavra "pneu" é visualmente animada segundo a transcrição fonética [pɪneʊ]. No contexto deste trabalho, os pseudo-encontros consonantais considerados são (DE MARTINO, 2005): /pt/, /pn/, /ps/, /bp/, /bb/, /bt/, /bt/, /bd/, /bk/, /bg/, /bm/, /bn/, /bf/, /bv/, /bs/, /bz/, /bʃ/, /bʒ/, /bɣ/, /tb/, /tk/, /tm/, /tn/, /ts/, /tz/, /tʃ/, /tʒ/, /dp/, /dk/, /dm/, /dn/, /dv/, /ds/, /dz/, /dʒ/, /kp/, /kb/, /kt/, /kd/, /km/, /kn/, /kf/, /ks/, /kz/, /gb/, /gd/, /gm/, /gn/, /gf/, /gs/, /mn/, /ms/, /ft/, /fn/, /vn/.

Da lista de pseudo-encontro consonantais apresentada, observa-se que a primeira posição de tais encontros pertence ao conjunto de fonemas {/p,b,t,d,k,g,m,f,v/}. Considerando-se os grupos homofemas da Tabela 4.1 e seus fones representantes, tal conjunto é simplificadamente representado pelos fones {/p,t,k,f/} (esta simplificação é adotada na Tabela 4.9).

Considerando-se os padrões silábicos apresentados nas tabelas 4.4, 4.5, 4.6 e 4.7 e os pseudo-encontros consonantais apresentados, é possível determinar os contextos fonéticos resultantes das diversas possibilidades de combinação de fones em uma palavra ou entre palavras, que envolvem a presença de segmentos consonantais. Assim, neste trabalho, os contextos considerados são expressos através dos seguintes padrões vogal-consoante:

- $\#CV$ - Fone consonantal precedido de silêncio e seguido de segmento vocálico (padrão silábico

Tab. 4.4: Padrões silábicos com um segmento consonantal no ataque e até um segmento consonantal em coda (adaptado de (DE MARTINO, 2005)).

Padrão Silábico	Exemplo	Segmento Consonantal
CV	**pa**ta	/p/
CVV	ju**deu**	/t/
CVVV	**qual**	/k/
CVC	**fis**co	/f/
CVVC	fa**lais**	/l/
CVVVC	i**guais**	/g/

Tab. 4.5: Padrões silábicos com dois segmentos consonantais no ataque e até um segmento consonantal em coda (adaptado de (DE MARTINO, 2005)).

Padrão Silábico	Exemplo	Encontro Consonantal
CCV	**pra**ta	/pɾ/
CCVV	**trei**no	/tɾ/
CCVC	te**clar**	/kl/
CCVVC	cate**drais**	/dɾ/

Tab. 4.6: Padrões silábicos sem ataque e um segmento consonantal em coda (adaptado de (DE MARTINO, 2005)).

Padrão Silábico	Exemplo	Segmento
VC	**es**pada	/s/
VVC	**aus**pício	/s/

Tab. 4.7: Padrão silábico com um segmento consonantal no ataque e dois em coda (adaptado de (DE MARTINO, 2005)).

Padrão Silábico	Exemplo	Segmento
CVCCC	**pers**pectiva	/p/

Tab. 4.8: Alguns exemplos de pseudo-encontros consonantais (adaptado de (DE MARTINO, 2005)).

Encontro	Exemplo
/pn/	**pn**eu
/tk/	vie**tc**ongue
/ft/	o**ft**álmico
/ks/	proli**x**o

CV da Tabela 4.4).

- $\#CCV$ - Encontro consonantal precedido de silêncio e seguido de segmento vocálico (padrão silábico CCV da Tabela 4.5).
- VCV - Resultante da combinação de quaisquer sílabas sem segmento consonantal na coda e padrão silábico CV da Tabela 4.4. Exemplo: $CVV + CV$ como em "**jei-to**" [1].
- $VCCV$ - Resultante das possíveis combinações:
  - sílaba sem segmento consonantal na coda combinada a sílaba com dois segmentos consonantais na posição de ataque. Exemplo: $CV + CCV$ como em "pa**to pre**to".
  - sílaba com um segmento consonantal na coda combinada a sílaba com um segmento consonantal na posição de ataque. Exemplo: $VC + CV$ como em "**es-pa**da".
- $VCCCV$ - Resultante das possíveis combinações:
  - sílaba com um segmento consonantal na coda combinada a sílaba com dois segmentos consonantais na posição de ataque. Exemplo: $VC + CCV$ como em "**es-pre**mer".
  - sílaba com dois segmentos consonantais na coda combinada a sílaba com um segmento consonantal na posição de ataque. Exemplo: $CVCC + CV$ como em "**pers-pe**ctiva".
- $VCCCCV$ - Resultante da combinação:
  - sílaba com dois segmentos consonantais na coda combinada a sílaba com dois segmentos consonantais na posição de ataque. Exemplo: $CVCC + CCV$ como em "**pers-tri**ção".

Considerando-se estes contextos, a Tabela 4.9 apresenta o seu mapeamento para contextos do tipo $\#CV$ e V_1CV_2. Segundo a abordagem adotada neste trabalho, este mapeamento é realizado através da análise do pentafone (sequência de 5 fones) que tem como fone central o fone a ser convertido em visema dependente de contexto. Assim, uma vez identificado um fone consonantal na sequência de fones proveniente da transcrição fonética, os dois fones vizinhos à esquerda do fone e os dois fones vizinhos à direita do fone são também analisados.

Um aspecto fundamental do mapeamento projetado para os contextos apresentados, são as seguintes características observadas para o Português do Brasil (DE MARTINO, 2005):

[1] O símbolo "-" é utilizado para realçar a divisão silábica em palavras.

- em sílabas com dois segmentos consonantais na posição de ataque, o segundo segmento consonantal é /ɾ/ ou /l/;

- sílabas com dois segmentos consonantais na posição de ataque em que o segundo segmento consonantal não é /ɾ/ ou /l/ contêm um pseudo-encontro consonantal;

- em sílabas com coda constituída por apenas um segmento consonantal, ocorrem apenas os fonemas /s/ e /r/;

- em sílabas com coda constituída por dois segmentos consonantais, ocorre apenas o encontro /ɾs/

Tais particularidades estão expressas na Tabela 4.9 e são informações essenciais para o mapeamento de contextos implementado. Nesta tabela, a primeira coluna apresenta o padrão vogal-consoante contemplado pelo mapeamento expresso na penúltima coluna da tabela. Na tabela são apresentadas 5 posições de cadeia de produção fonética. O segmento sonoro da posição 1 é produzido imediatamente antes do segmento da posição 2, e assim sucessivamente. O fone de interesse para o mapeamento em um dado instante encontra-se na posição central 3. Na tabela adotou-se as seguintes convenções:

- o segundo fone à esquerda do fone central é identificado pelo índice 1 (Posição 1);

- o primeiro fone à esquerda do fone central é identificado pelo índice 2 (Posição 2);

- o fone central é identificado pelo índice 3 (Posição 3);

- o primeiro fone à direita do fone central é identificado pelo índice 4 (Posição 4);

- o segundo fone à direita do fone central é identificado pelo índice 5 (Posição 5);

- o símbolo "#" indica silêncio;

- o símbolo V indica segmento vocálico;

- o símbolo C indica segmento consonantal;

- a notação V_i indica o segmento vocálico na posição i;

- o símbolo "*" (asterisco) indica que não importa o tipo de fone;

- a notação [i] indica o segmento sonoro [i].

Tab. 4.9: Mapeamento de contextos fonéticos

Contexto Considerado	Posicionamento					Mapeamento	Exemplo
	Segundo Fone Esquerda 1	Primeiro Fone Esquerda 2	Fone Central 3	Primeiro Fone Direita 4	Segundo Fone Direita 5		
$\#CV$	*	#	C_3	V_4	*	$\#C_3V_4$	**Pa**ta
$\#CCV$	*	#	C_3	$C_4 \in${[l,ɾ]}	V_5	$\#C_3V_5$	**Pla**ca
	*	#	$C_3 \in${[p,t,f,k]}	$C_4 \notin${[l,ɾ]}	V_5	$\#C_3$[i]	**Pne**u
	#	C_2	$C_3 \in${[l,ɾ]}	V_4	*	$V_4C_3V_4$	**pLa**ca
	#	$C_2 \in${[p,t,f,k]}	$C_3 \notin${[l,ɾ]}	V_4	*	[i]C_3V_4	**pNe**u
VCV	*	V_2	C_3	V_4	*	$V_2C_3V_4$	**jeiTo**
$VCCV$	*	V_2	C_3	$C_4 \in${[l,ɾ]}	V_5	$V_2C_3V_5$	**aPli**cação
	*	V_2	$C_3 \in${[p,t,f,k]}	$C_4 \notin${[l,ɾ]}	V_5	V_2C_3[i]	**aPné**ia
	*	V_2	$C_3 \in${[s,ɾ]}	C_4	V_5	$V_2C_3V_2$	**aRti**sta
	V_1	C_2	$C_3 \in${[l,ɾ]}	V_4	*	$V_4C_3V_4$	**apLi**cação
	V_1	$C_2 \in${[p,t,f,k]}	$C_3 \notin${[l,ɾ]}	V_4	*	[i]C_3V_4	**apNé**ia
	V_1	$C_2 \in${[s,ɾ]}	C_3	V_4	*	$V_1C_3V_4$	**arTi**sta
$VCCCV$	*	V_2	$C_3 \in${[s,ɾ]}	C_4	C_5	$V_2C_3V_2$	**aRtri**te
	*	V_2	$C_3 =$ɾ	$C_4 =$s	C_5	$V_2C_3V_2$	p**eRspe**ctiva
	V_1	$C_2 \in${[s,ɾ]}	C_3	$C_4 \in${[l,ɾ]}	V_5	$V_1C_3V_5$	**arTri**te
	V_1	$C_2 \in${[s,ɾ]}	$C_3 \in${[p,t,f,k]}	$C_4 \notin${[l,ɾ]}	V_5	V_1C_3[i]	**os Pne**us
	V_1	$C_2 =$ɾ	$C_3 =$s	C_4	V_5	$V_1C_3V_1$	p**erSpe**ctiva
	$C_1 \in${[s,ɾ]}	C_2	$C_3 \in${[l,ɾ]}	V_4	*	$V_4C_3V_4$	**artRi**te
	$C_1 \in${[s,ɾ]}	$C_2 \in${[p,t,f,k]}	$C_3 \notin${[l,ɾ]}	V_4	*	$[i]C_3V_4$	**os pNe**us
	$C_1 =$ɾ	$C_2 =$s	C_3	V_4	*	$\#C_3V_4$	p**ersPe**ctiva
$VCCCCV$	*	V_2	C_3=ɾ	C_4=s	C_5	$V_2C_3V_2$	p**eRscru**tar
	V_1	C_2=ɾ	C_3=s	C_4	C_5	$V_1C_3V_1$	p**erScru**tar
	C_1=ɾ	C_2=s	C_3	$C_4 \in${[l,ɾ]}	V_5	$\#C_3V_5$	p**ersCru**tar
	C_1=s	C_2	$C_3 \in${[l,ɾ]}	V_4	*	$V_4C_3V_4$	p**erscRu**tar

4.4 Síntese dos Quadros de Animação

Após o processamento das informações temporais fornecidas pela transcrição fonética temporizada (Seção 4.2) e a conversão de fones em visemas (Secao 4.3), obtém-se uma linha temporal de animação em que os instantes correspondentes aos alvos articulatórios são associados a visemas que representam as poses-chave, ou visemas-chave, da animação a ser sintetizada (vide Figura 4.4). É importante notar que os instantes correspondentes aos alvos articulatórios podem ou não coincidir com a ocorrência de quadros da animação, não afetando a maneira como os quadros intermediários aos visemas-chave são sintetizados.

Os quadros de animação são gerados considerando-se a taxa de reprodução de 30 quadros por segundo e atribui-se o primeiro quadro ao instante zero da animação.

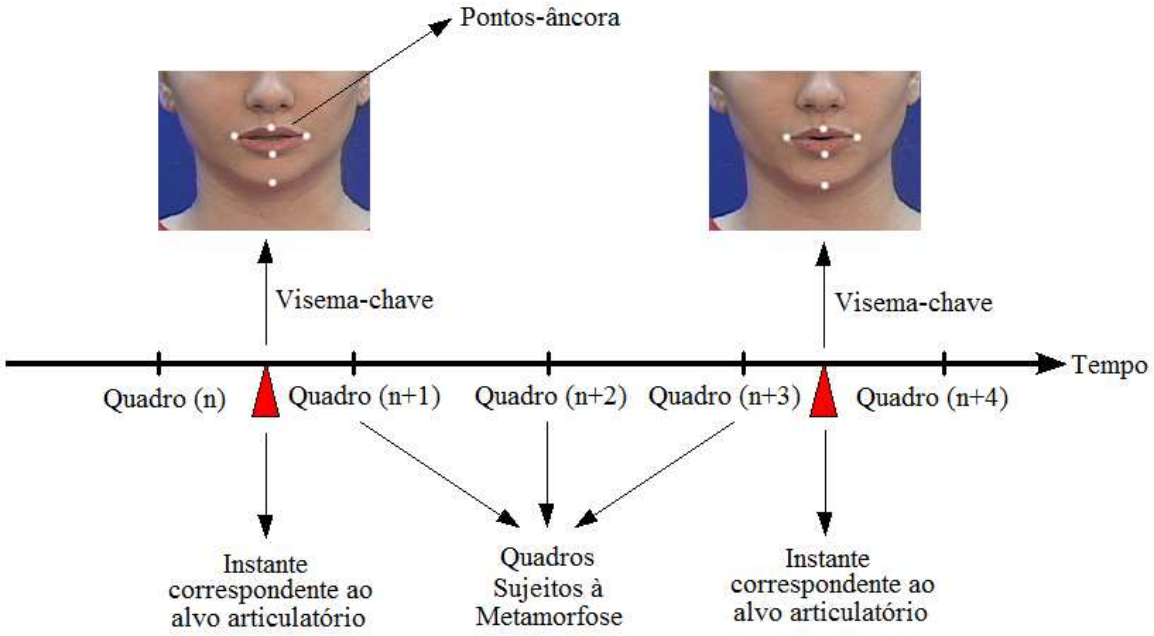


Fig. 4.4: Linha temporal de animação resultante do processamento das informações da transcrição fonética temporizada

O processo de síntese da animação consiste, fundamentalmente, na geração dos quadros intermediários entre visemas-chave subsequentes, sintetizados através do processo de metamorfose entre imagens.

A metamorfose entre imagens é implementada a partir de sucessivas transformações espaciais, ou *warping*, em que uma imagem de origem é distorcida gradualmente, através de quadros intermediários, em direção a uma imagem de destino. Analogamente, o *warping* é aplicado à imagem de destino distorcendo-a gradualmente em direção à imagem de origem. A metamorfose é concluída combinando-se os resultados das distorções espaciais em sentidos opostos através de uma operação de dissolução cruzada (*cross-dissolve*).

A Figura 4.5 ilustra simplificadamente este processo. Na figura mostra-se, no canto superior esquerdo, um exemplo de visema de origem cuja imagem deve ser transicionada de maneira suave

e contínua, no visema de destino apresentado no canto inferior direito da figura. Neste exemplo, a transição é realizada através de apenas dois quadros intermediários. Na primeira linha da figura, mostram-se as duas imagens resultantes do *warping* do visema de origem em direção ao visema de destino, no sentido direto da transformação espacial. Da mesma maneira, na última linha da figura, observa-se a distorção aplicada ao visema de destino na direção do visema de origem, no sentido reverso da transformação espacial. Finalmente, a linha intermediária da figura apresenta os visemas finais sintetizados, obtidos a partir da dissolução cruzada entre as imagens da primeira e última linhas da figura, na mesma coluna.

O objetivo desta seção é descrever como o algoritmo de metamorfose foi implementado no contexto deste trabalho. Uma visão mais abrangente das etapas de um processo de metamorfose entre imagens e diferentes abordagens são descritas com maior profundidade, por exemplo, em (WOLBERG, 1998).

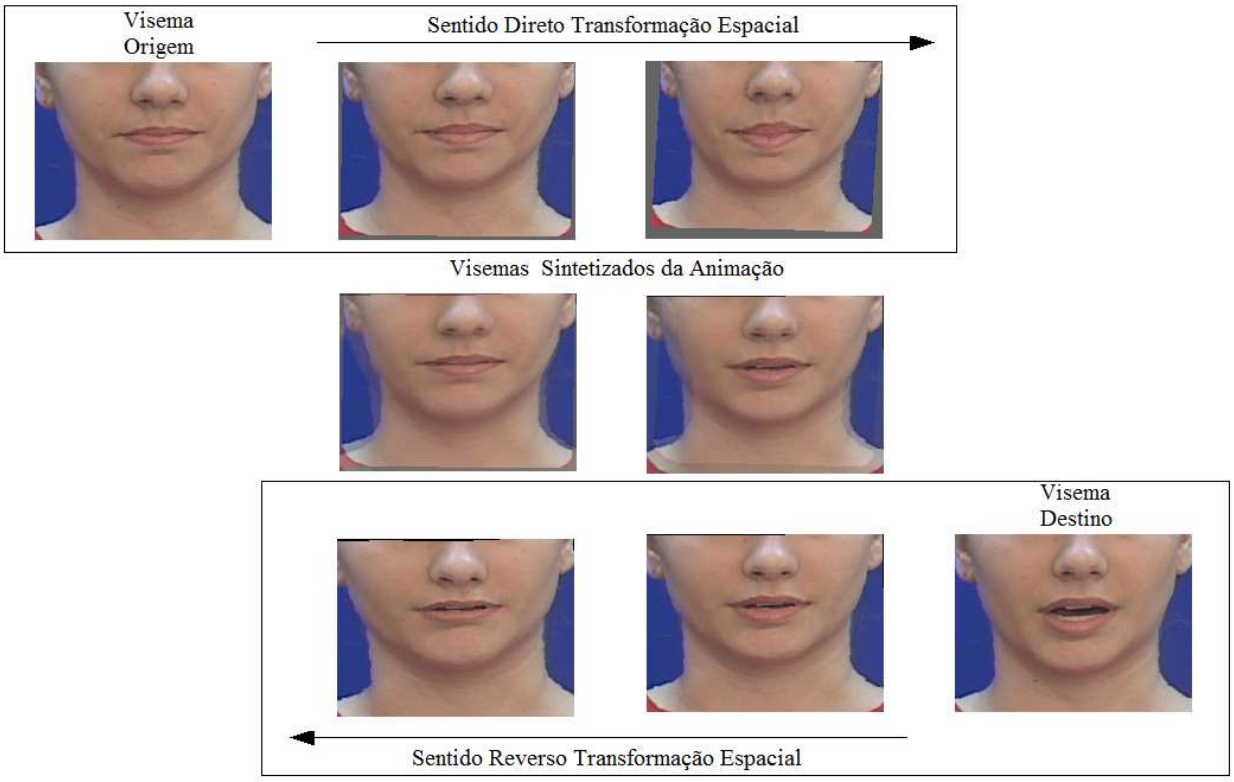


Fig. 4.5: Processo de metamorfose entre visemas-chave

Trajetória de Transição entre Visemas-Chave

A geração dos quadros intermediários é guiada pelos pontos-âncora, utilizados pelo processo de *warping* e detectados durante o processo de construção da base de imagens (Seção 3.5.3). A Figura 4.6 apresenta os 5 pontos-âncora detectados para um dos visemas da base de imagens.

A determinação dos pontos-âncora para os quadros intermediários é realizada a partir da definição da trajetória das coordenadas x e y destes pontos em função do tempo, tomando-se como referência os

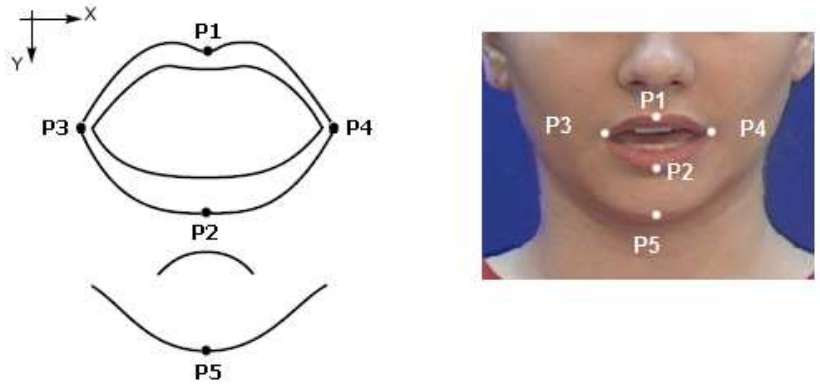


Fig. 4.6: Pontos-âncora associados aos visemas-chave

pontos-âncora dos visemas-chave de origem e destino. Uma vez que a definição desta trajetória interfere diretamente na percepção visual da dinâmica dos movimentos articulatórios, optou-se por utilizar uma modelagem baseada em uma curva de interpolação não-linear. Tal abordagem implementa uma função de transição cujas características procuram modelar mais adequadamente a complexa e variável dinâmica de transição entre alvos articulatórios e representa uma alternativa à transições lineares implementadas por sistemas como (EZZAT; POGGIO, 1998), (GOYAL; KAPOOR; KALRA, 2000), (FARUQUIE et al., 2001).

Para isso, adotou-se uma curva de interpolação paramétrica de Hermite (FOLEY, 1990) que possui características de continuidade geométrica G^0 entre os vários quadros de uma sequência e garantia de derivada temporal igual a zero nos instantes de realização das poses-chave. Abordagem semelhante foi adotada em (DE MARTINO, 2005), onde também são apresentadas algumas comparações entre transições de alvos articulatórios medidas em laboratório e a curva de transição gerada através da curva cúbica de Hermite adotada.

A curva de interpolação é obtida através das equações:

$$\begin{bmatrix} x_i(t) \\ y_i(t) \end{bmatrix} = \begin{bmatrix} Ox_i & Dx_i \\ Oy_i & Dy_i \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} 0 \leq t \leq 1$$

- $x_i(t)$ e $y_i(t)$ representam as funções de Hermite para as coordenadas (x, y) do i-ésimo ponto-âncora de interesse, onde $i = 1...5$;
- Ox_i e Oy_i são as coordenadas x e y do ponto-âncora de interesse para a imagem origem;
- Dx_i e Dy_i são as coordenadas x e y do ponto-âncora de interesse para a imagem destino;
- *t* é a variável independente da representação paramétrica normalizada em relação ao intervalo de tempo entre os visemas-chave origem e destino.

Transformação Espacial de Imagens Utilizando Funções de Base Radial

Uma vez determinados os pontos-âncora dos quadros intermediários entre visemas-chave, segue-se o processo de distorção dos visemas-chave origem e destino nos sentidos direto e reverso da metamorfose, como exemplificado pela Figura 4.5.

Considerando-se o pequeno número de pontos-âncora considerados, o processo de *warping* aplicado aos visemas-chave seguiu a mesma abordagem implementada durante o processo de registro das imagens da base e apresentada, em detalhes, na Seção 3.5.2. Conforme discutido anteriormente, a transformação geométrica espacial de imagens considerando-se um pequeno número de pontos de controle representa um problema de interpolação de dados esparsos no espaço. Considerando-se ainda a natureza das imagens a serem distorcidas, adotou-se funções de base radial (RBF - *Radial Basis Functions*) como estratégia de *warping* dos visemas. Nesta abordagem a função de interpolação é uma combinação linear de funções radiais, em que os *pixels* equidistantes de um mesmo ponto-âncora são influenciados de maneira idêntica por este ponto durante o processo de transformação.

A Seção 3.5.2 e o Apêndice B apresentam a formulação empregada para a transformação dos visemas-chave.

Face-base

Após a síntese dos quadros entre visemas-chave, obtém-se uma sequência de visemas cujas imagens são restritas à região da face composta por lábios e queixo. A próxima etapa no processo de síntese da animação final, é a fusão destas imagens a uma imagem de referência, ou face-base.

A combinação de um visema à face-base é realizada a partir da aplicação de uma máscara de transparência ao visema e da combinação desta imagem à face-base, já previamente processada.

A máscara de transparência é definida como uma imagem *bitmap*, $\alpha(x, y)$, de mesmas dimensões das imagens originalmente extraídas da trilha de vídeo do corpus audiovisual (720 x 486 *pixels*), com valores de *pixels* variando entre zero (*pixels* pretos) e um (*pixels* brancos).

O projeto da máscara de transparência foi realizado através da definição manual de dois contornos na imagem de referência. Inicialmente, um contorno, denominado contorno interno **CI**, foi desenhado ao redor da região facial em que se observa a maior influência da movimentação articulatória visível, constituída por lábios, bochechas e queixo. Em seguida, definiu-se um contorno externo **CE**, externo à região delimitada pelo contorno **CI**, que define uma região de fronteira entre a face-base e a região contemplada pelos visemas da base de imagens. A Figura 4.7(a) ilustra o projeto da máscara através dos contornos **CI** e **CE**.

Após a definição de tais contornos, criou-se uma imagem *bitmap* de mesmas dimensões da imagem de referência, em que se atribuiu valor "0" (totalmente transparentes) a todos os *pixels* localizados externamente ao contorno **CE** e valor "1" (totalmente opacos) a todos os *pixels* localizados internamente ao contorno **CI**. Na região entre os dois contornos, a máscara é parcialmente transparente, com valores de *pixel* entre "0" e "1", em função da distância do pixel aos contornos externo e interno. Seja:

- P um ponto pertencente ao domínio da imagem, com coordenadas (x, y);
- A o conjunto de pontos internos ao contorno CI;
- B o conjunto de pontos internos ao contorno CE;

- $dist(C, P)$ a menor distância euclidiana entre o ponto P e o contorno C.

Tem-se:

$$\alpha_P = \begin{cases} 1, \{P : P \in A\} \\ 0, \{P : P \notin B\} \\ \frac{dist(CE,P)}{dist(CE,P)+dist(CI,P)}, \{P : P \in B \text{ e } P \notin A\} \end{cases}$$

Na Figura 4.7(b) tem-se uma ilustração do resultado da máscara gerada.

(a) Contornos utilizados para gerar máscara de transparência (b) Máscara de transparência

Fig. 4.7: Projeto da máscara de transparência utilizada para fundir um visema a uma face-base

A fusão da face-base ao visema, através da aplicação da máscara de transparência pode ser expressa da seguinte maneira:

$$I(x, y) = \alpha_p \cdot f(x, y) + (1 - \alpha_p) \cdot g(x, y) \tag{4.2}$$

onde:

- $f(x, y)$ é a imagem da face-base;
- $g(x, y)$ é o visema sintetizado da animação;
- $I(x, y)$ é a imagem final combinando-se a face-base e o visema.

Na Figura 4.8, a Figura 4.8(a) mostra o resultado da aplicação da máscara de transparência à imagem da face-base. A Figura 4.8(b) mostra um exemplo de visema sintetizado e o resultado da aplicação da máscara complementar a este visema, conforme equação 4.2. Finalmente, a Figura 4.8(c) apresenta o resultado final da fusão do visema à face-base, formando a imagem correspondente a um quadro da animação final.

4.5 Comentários Finais

Neste capítulo, detalhou-se o processo de síntese da animação implementado por este trabalho. A partir do fornecimento da transcrição fonética temporizada da fala a ser visualmente animada, o sistema obtém as informações temporais e de segmentos da fala necessárias para a síntese dos quadros da animação através da estratégia de metamorfose entre visemas-chave.

(a) Face-base após aplicação da máscara de transparência

(b) Visema a ser fundido com a face-base e o resultado após aplicação da máscara complementar

(c) Quadro final da animação

Fig. 4.8: Fusão do visema sintetizado da animação à face-base, gerando o quadro final da animação (Equação 4.2).

Uma parcela essencial da modelagem da movimentação articulatória visível da fala é implementada através da etapa de conversão de fones em visemas, detalhadamente abordada na Seção 4.3. Nesta etapa, o sistema se baseia em uma tabela de mapeamento de fones e contextos fonéticos para os visemas dependentes de contexto considerados neste trabalho (vide Capítulo 3). O mapeamento implementado foi projetado a partir da análise das características dos padrões silábicos existentes para o Português do Brasil. Tal análise pode ser semelhantemente realizada para outras línguas.

A partir da identificação dos visemas-chave, a Seção 4.4 detalhou os algoritmos de processamento de imagens empregados para a síntese dos quadros da animação. Nesta etapa, importantes aspectos da implementação adotada neste trabalho podem ser destacados.

Em primeiro lugar, o sistema implementado opta por processar a animação apenas na região dos lábios e queixo, região efetivamente influenciada pela movimentação articulatória visível da fala. Esta abordagem reflete, inicialmente, o foco deste trabalho em propor uma alternativa à modelagem da movimentação articulatória para sistemas de animação facial 2D. Adicionalmente, entende-se que tal abordagem possibilita trabalhos futuros em que a modelagem da movimentação de diferentes partes faciais (como olhos e sobrancelhas) e da cabeça possa ser realizada separadamente e de maneira modular, implementando-se uma arquitetura flexível de animação facial 2D em que novos recursos podem ser acoplados, visando atribuir maior vídeo-realismo à animação final. Dessa maneira, esta abordagem se opõe a abordagens em que a imagem da face como um todo é tratada e que tipicamente acarretam em bases de imagens extensas quando o objetivo é armazenar imagens com diferentes expressões faciais ou movimentações da cabeça combinadas a diferentes visemas.

Em segundo lugar, uma característica importante do processo de metamorfose entre visemas implementado é a utilização de funções de base radial como funções de transformação espacial e a utilização de apenas 5 pontos-âncora que guiam o processo de metamorfose. Esta implementação, além de produzir bons resultados visuais nas imagens sintetizadas, pode ser considerada computacionalmente simples, uma vez que envolve a solução de sistemas determinados de equações lineares e operações algébricas não complexas. Esta característica permite que a solução seja embarcada em sistemas com capacidades de processamento e memória dinâmica limitadas.

Adicionalmente, o sistema adota uma curva de transição temporal não-linear entre visemas-chave. A abordagem adotada, também implementada em (DE MARTINO, 2005), permite uma modelagem mais próxima da dinâmica não-linear observada na transição entre alvos-articulatórios correspondentes a segmentos da fala.

Capítulo 5

Implementação Piloto do Sistema de Animação Facial 2D

5.1 Introdução

Os capítulos 3 e 4 deste trabalho descrevem o processo de síntese de animação facial 2D baseado na metamorfose entre visemas dependentes de contexto, a partir de uma base de visemas de apenas 34 imagens. Neste capítulo, descrevem-se os principais aspectos da implementação piloto do sistema de animação facial 2D que emprega os conceitos e metodologias apresentados nos capítulos anteriores. Esta implementação permite a validação e avaliação da abordagem de síntese proposta neste trabalho. Com o sistema piloto foram geradas as animações utilizadas durante a avaliação do sistema, discutida no Capítulo 6.

O sistema piloto implementa uma *talking head* a partir da integração entre o sistema de síntese de animação facial 2D desenvolvido no contexto deste trabalho e o conversor texto-fala "CPqD Texto Fala". A Figura 5.1 apresenta o diagrama de blocos do sistema piloto de síntese da animação facial sincronizada com a fala.

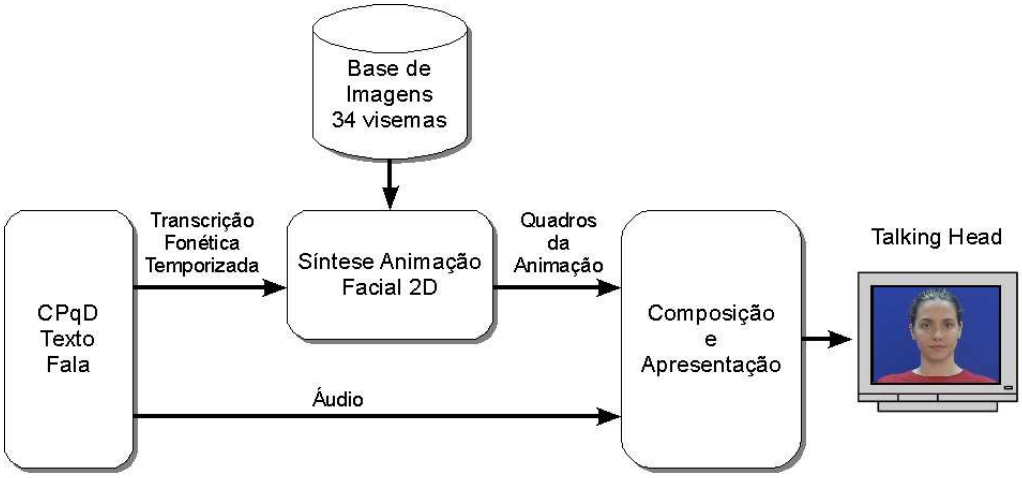


Fig. 5.1: Implementação piloto do sistema de animação facial 2D sincronizada com a fala

O "CPqD Texto Fala" é um sistema de síntese da voz humana desenvolvido e comercializada pela Fundação CPqD, centro de pesquisa e desenvolvimento em telecomunicações. A partir de uma entrada textual, o "CPqD Texto Fala" gera o sinal de áudio correspondente à fala. No sistema piloto, além do áudio, o "CPqD Texto Fala" fornece também a transcrição fonética temporizada da fala sintetizada. A Seção 5.2 descreve a utilização do conversor "CPqD Texto Fala" na implementação piloto.

A Seção 5.3 apresenta os principais aspectos de desenvolvimento do software de síntese da animação facial 2D, fornecendo uma visão geral sobre a estrutura do software e relacionando seus módulos principais aos conceitos apresentados nos capítulos anteriores.

Ao final deste capítulo, na Seção 5.4, realiza-se uma breve análise sobre a performance do sistema implementado considerando-se suas principais características.

5.2 Conversor Texto-Fala "CPqD Texto Fala"

"CPqD Texto Fala" é o nome do sistema comercial de conversão texto-fala, ou sistema TTS (*text-to-speech*), desenvolvido pela Fundação CPqD. Este sistema é caracterizado por um software que tem a capacidade de sintetizar sinais de fala a partir de texto escrito em Português e expresso em código ASCII (*American Standard Code for Information Interchange*).

O sistema "CPqD Texto Fala" consiste em uma biblioteca de ligação dinâmica (*DLL - dynamic link library*) que contém funções para inicialização, uso e finalização do mecanismo de conversão texto-fala. Esta biblioteca pode ser integrada a qualquer tipo de aplicação, independentemente da linguagem de programação utilizada.

A versão da biblioteca utilizada neste trabalho gera voz feminina, em arquivos WAV, no formato: PCM (*Pulse Code Modulation*) linear (16 bits/amostra) em 16 kHz, com cabeçalho. Além da versão utilizada neste trabalho, existem versões da biblioteca capazes de gerar voz masculina ou feminina em cada um dos seguintes formatos: PCM linear (8 ou 16 bits/amostra), PCM lei-A (8 bits/amostra) e PCM lei-μ (8 bits/amostra), em 8 ou 16 kHz, com ou sem cabeçalho WAV (CPQD, 2007).

Em sua implementação, o "CPqD Texto Fala" é constituído de 4 módulos principais (COSTA; DE MARTINO; NAGLE, 2008):

- **Módulo de pré-processamento:** transforma o texto de entrada em um fluxo de palavras e símbolos de pontuação; números, símbolos especiais e abreviações, por exemplo, são convertidos em palavras escritas por extenso.

- **Módulo de conversão ortográfico-fonética:** converte a sequência de palavras obtidas na saída do módulo de pré-processamento na sequência de unidades fonéticas correspondentes à sentença que será sintetizada.

- **Módulo de síntese:** gera um sinal de fala sintética a partir da sequência de unidades fonéticas gerada pelos módulos anteriores e de uma base de dados de fala natural acoplada ao sistema. O sintetizador seleciona unidades acústicas apropriadas na sua base de dados e as processa, inserindo ritmo e entonação de modo a obter fala sintetizada de alta qualidade.

- **Módulo de codificação:** converte a saída gerada para o formato de áudio desejado para a aplicação.

A transcrição fonética temporizada, fornecida como entrada para o processo síntese da animação facial, é composta pela sequência de unidades fonéticas utilizadas na geração do sinal de fala associadas às respectivas durações. Ela é disponibilizada através de um arquivo texto, como resultado intermediário, pelo módulo de síntese.

O texto a seguir exemplifica como as informações da transcrição fonética temporizada são disponibilizadas pelo "CPqD Texto Fala". A transcrição apresentada foi obtida a partir da locução da frase: "Não pode haver tréguas na guerra contra a malária."

```
0: 680 ##
1: 1048 nn
2: 3464 'a~
3: 5686 uu
4: 8694 pp
5: 10538 'OO
6: 14922 dZ
7: 17258 ii
8: 18146 aa
9: 19096 vv
10: 20784 'ee
11: 23782 rr
.
.
.
31: 72064 mm
32: 73692 aa
33: 76562 ll
34: 78538 'AA
35: 83138 r^
36: 84944 ii
37: 86682 aa
38: 89546 ##
39: 95270
```

É possível perceber a existência de 3 campos distintos em cada linha do arquivo de transcrição fonética:

- o primeiro campo (delimitado pelo símbolo ":") identifica a ordem do fone na sequência de fones da locução;
- o segundo campo indica a posição em *bytes* no arquivo WAV, do início do fone considerado;
- o terceiro campo indica o fone considerado segundo simbologia específica adotada pelo "CPqD Texto Fala".

As tabelas 5.1 e 5.2 apresentam a correspondência entre os símbolos adotados pelo "CPqD Texto Fala" e os fones considerados neste trabalho, tendo-se como referência o Alfabeto Fonético Internacional (IPA - *International Phonetic Alphabet*) (INTERNATIONAL PHONETIC ASSOCIATION, 1999).

Em particular, o símbolo "*##*" não representa um fone, mas é um indicador de silêncio (vide a primeira linha da transcrição fonética temporizada apresentada).

Considere, por exemplo, a seguinte linha da transcrição apresentada:

Tab. 5.1: Correspondência entre notação utilizada na transcrição fonética temporizada fornecida pelo "CPqD Texto Fala" e fones do Alfabeto Fonético Internacional (IPA - *International Phonetic Alphabet*) (INTERNATIONAL PHONETIC ASSOCIATION, 1999) - Segmentos Consonantais.

Símbolo "CPqD Texto Fala"	Símbolo IPA
pp	[p]
bb	[b]
mm	[m]
ff	[f]
vv	[v]
tt, tS	[t]
dd, dZ	[d]
nn	[n]
kk	[k]
gg	[g]
ss	[s]
SS	[ʃ]
zz	[z]
ll	[l]
lˆ	[ʎ]
nˆ	[ɲ]
RR, rr	[ɣ]
rˆ	[ɾ]

Tab. 5.2: Correspondência entre notação utilizada na transcrição fonética temporizada fornecida pelo "CPqD Texto Fala" e fones do Alfabeto Fonético Internacional (IPA - *International Phonetic Alphabet*) (INTERNATIONAL PHONETIC ASSOCIATION, 1999) - Segmentos Vocálicos.

Símbolo "CPqD Texto Fala"	Símbolo IPA
II	[i]
i~	[ĩ]
ee	[e]
e~	[ẽ]
EE	[ɛ]
AA	[a]
a~	[ɐ̃]
OO	[ɔ]
oo	[o]
o~	[õ]
UU	[u]
u~	[ũ]
ii	[ɪ]
aa	[ɐ]
uu	[ʊ]

```
\textit{37: 86682 aa}
```

Nesta linha, o símbolo "*aa*" representa a locução do fone [ɐ]. Este fone é o trigésimo sétimo fone da sequência apresentada e sua fronteira inicial [ɐ] está posicionada a *86682* bytes do início do arquivo WAV correspondente.

Considerando-se a versão de biblioteca utilizada neste trabalho, o arquivo WAV gerado possui cabeçalho de 44 bytes, 16 bits/amostra e taxa de amostragem de 16 kHz. Desta maneira, a informação temporal da fronteira de início de cada fone da fala sintetizada pelo conversor "CPqD Texto Fala" pode ser obtida através da seguinte fórmula:

$$f_s = \frac{(f_{bytes} - c_{WAV})}{2 \times freq} = \frac{(f_{bytes} - 44)}{2 \times 16000} \tag{5.1}$$

onde:

- f_{bytes} - fronteira inicial do fone, informada no arquivo de transcrição fonética fornecido pelo sistema "CPqD Texto Fala", dada em bytes;
- c_{WAV} - tamanho do cabeçalho do arquivo WAV em bytes;
- $freq$ - frequência de amostragem do arquivo WAV em hertz;
- f_s - fronteira inicial do fone, dada em segundos.

5.3 Software de síntese da animação facial 2D

O desenvolvimento do software de síntese da animação facial 2D foi realizado segundo algumas diretrizes básicas que visaram:

- a adequada implementação dos algoritmos relacionados ao processo de síntese da animação;
- possibilitar a fácil expansão e/ou agregação de novas funcionalidades do sistema decorrentes de trabalhos futuros;
- possibilitar que a implementação disponibilizada fosse facilmente compreendida, suportada e replicável em diferentes plataformas de hardware.

Seguindo tais diretrizes, o software foi desenvolvido utilizando-se a linguagem de programação C++ segundo uma estrutura orientada a objetos. O sistema foi desenvolvido e testado a partir do ambiente de desenvolvimento integrado Dev-C++ utilizando o compilador GCC (*Gnu Compiler Collection*), em plataforma Windows XP. Para operações especializadas (como processamento de texto ou operações de processamento de imagens), o sistema utilizou bibliotecas livres e de código aberto, consideradas bem documentadas e robustamente suportadas.

O fluxograma da Figura 5.2 fornece uma visão simplificada do processamento implementado pelo software que sintetiza a animação facial a partir da transcrição fonética temporizada fornecida pelo conversor texto-fala.

A seguir, cada bloco do diagrama é brevemente descrito, mapeando-se as etapas do processo de síntese da animação facial 2D aos conceitos apresentados nos capítulos anteriores e apontando-se os aspectos considerados mais relevantes relacionados à implementação.

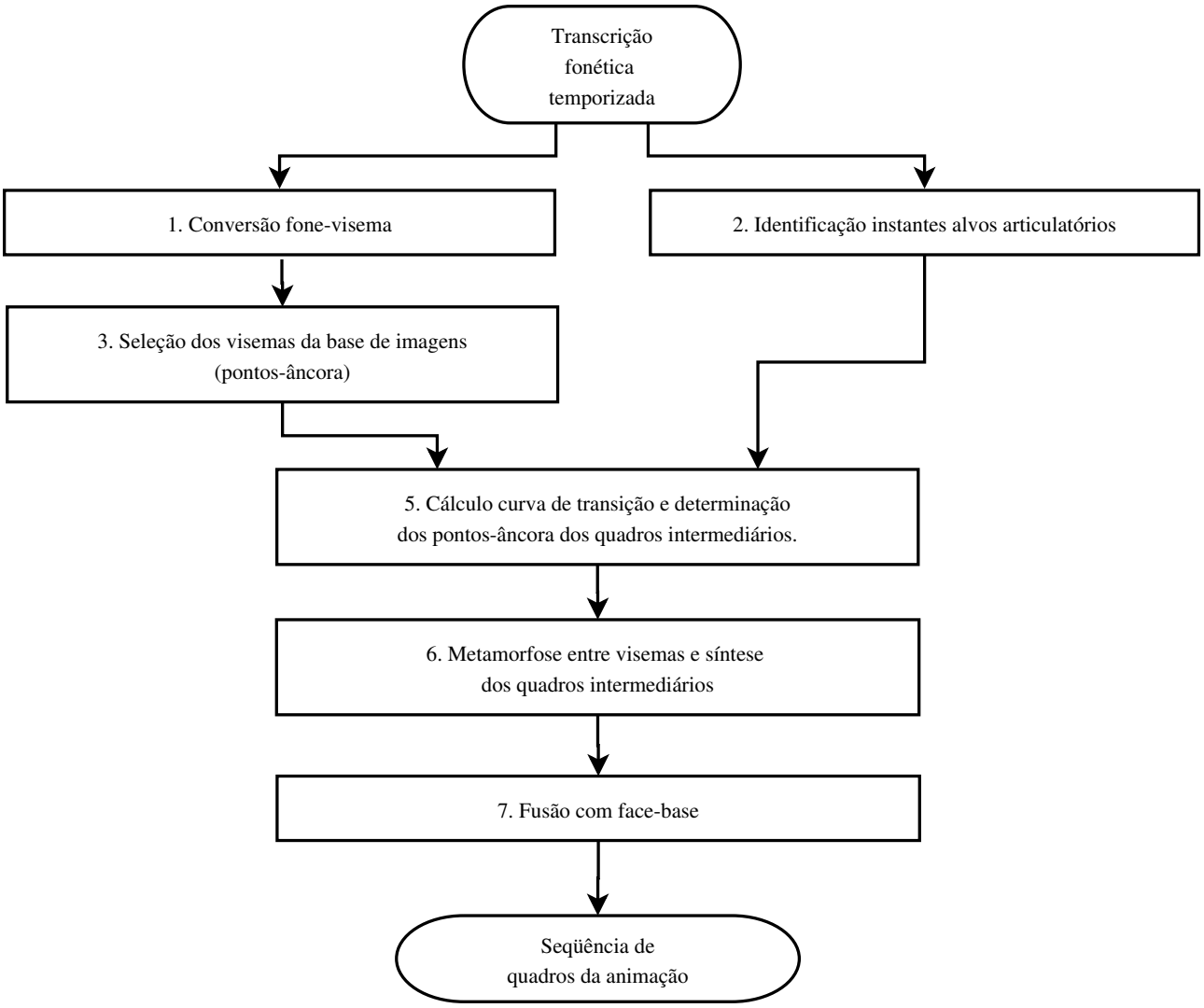


Fig. 5.2: Diagrama de blocos da implementação piloto

1. **Conversão fone-visema**

   Conforme discutido na Seção 5.2, a transcrição fonética temporizada fornecida pelo "CPqD Texto Fala" é disponibilizada através de um arquivo texto. O arquivo texto fornecido é processado, possibilitando a extração das informações de sequência de fones e fronteiras iniciais de cada fone. Nesta etapa, uma das operações realizadas é a conversão dos símbolos de fones utilizados pelo "CPqD Texto Fala" na notação equivalente que permite realizar a conversão de fones em visemas, descrita no Capítulo 4 (Seção 4.3).

2. **Identificação instantes alvos articulatórios**

   O processamento do arquivo da transcrição fonética temporizada permite a determinação dos instantes associados aos alvos articulatórios, conforme processo descrito na Seção 4.2. Como descrito nessa seção, a partir das informações temporais fornecidas pela transcrição fonética temporizada, é possível determinar o número de quadros da animação considerando a apresentação dos quadros a uma taxa de 30 quadros por segundo.

3. **Seleção dos visemas da base de imagens**

   O Capítulo 3 apresentou o processo de construção da base de imagens, constituída de 34 imagens correspondentes a visemas mais informações de rótulo associadas (vide Seção 3.6). As informações de rótulo foram organizadas em um arquivo texto CSV (*Comma-separated values*), formato de arquivo que armazena dados tabelados, amplamente utilizado para conversão e exportação de dados entre planilhas de dados e sistemas de gerenciamento de base de dados (SHAFRANOVICH, 2005). O arquivo texto CSV utilizado nesta implementação caracteriza uma pequena base de dados, a partir da qual podem ser recuperados os registros associados a cada visema-chave da animação a ser sintetizada. Cada registro desta base de dados contém as seguintes informações:

   - caminho para o arquivo da imagem;
   - identificador do fragmento de vídeo do qual a imagem foi extraída;
   - número do quadro correspondente à imagem no fragmento de vídeo do qual a imagem foi extraída;
   - identificador do visema de acordo com as tabelas 3.3 e 3.4;
   - pentafone (sequência de 5 fones) que indica o contexto fonético no qual a imagem foi capturada incluindo: 2 fones adjacentes à esquerda do fone central, fone central, 2 fones adjacentes à direita do fone central;
   - coordenadas x e y dos 5 pontos-âncora medidos para o visema (Figura 3.8).

   A recuperação destes dados é o resultado da seleção dos visemas da base de imagens.

4. **Cálculo da curva de transição e determinação dos pontos-âncora dos quadros intermediários**

   A partir dos pontos-âncora dos visemas-chave e a determinação dos instantes associados aos quadros intermediários entre dois visemas-chave subsequentes, é possível implementar o cálculo dos pontos-âncora dos visemas a serem sintetizados nos quadros intermediários. Esta

operação é realizada através do cálculo da curva de transição para cada um dos pontos-âncora, conforme processo descrito na Seção 4.4.

5. **Metamorfose entre visemas e síntese dos quadros intermediários**

   Prosseguindo-se com a síntese dos quadros intermediários entre visemas-chave, o próximo passo implementado pelo aplicativo é a metamorfose entre os visemas-chave. Nesta etapa, a biblioteca de visão computacional OpenCV (BRADSKI; KAEHLER, 2008) foi utilizada para implementação das operações de *warping* e dissolução cruzada descritas na Seção 4.4. A opção pela implementação através de funções desta biblioteca foi realizada visando a manipulação de imagens de maneira computacionalmente eficiente, a partir de uma plataforma de processamento de imagens especializada e bem documentada. A Seção 5.3.1 deste capítulo fornece maiores detalhes sobre esta biblioteca.

6. **Fusão com face-base**

   Conforme descrito na Seção 4.4, após a síntese dos visemas associados à animação, o último passo do processo corresponde à fusão dos visemas sintetizados a uma face-base. Mais uma vez, a biblioteca de visão computacional OpenCV foi utilizada para implementar a operação de fusão entre estas imagens. A saída do processo de síntese implementado é caracterizada por um conjunto de imagens, tendo-se como referência a taxa de reprodução de 30 quadros por segundo, com o primeiro quadro associado ao instante zero da duração da animação.

O aplicativo desenvolvido produz como saída uma sequência de imagens correspondentes aos quadros da animação sintetizada. A partir da sequência de imagens geradas e o aúdio da fala sintetizado pelo conversor texto-fala, segue-se o processo de composição e apresentação da animação final.

O processo de composição e apresentação pode ser realizado em tempo de síntese ou pode ser o resultado de um processo *offline* de mixagem entre áudio e quadros da animação. Na primeira abordagem, após a síntese da animação, os quadros gerados são imediatamente apresentados de maneira síncrona com o áudio da fala correspondente, numa taxa de reprodução de 30 quadros por segundo. Já o processo de mixagem entre áudio e vídeo permite a geração de um vídeo de animação que pode ser adaptado a diferentes formatos de reprodução visando diferentes plataformas de destino, permitindo também a aplicação de algoritmos de compressão de vídeo.

A mixagem de áudio e vídeo pode ser realizada através de ferramentas especializadas que permitem diversos níveis de edição do arquivo de vídeo gerado. No contexto deste trabalho, duas ferramentas foram utilizadas, e são citadas como exemplos de ferramentas de edição:

- **FFmpeg:** ferramenta de gravação, conversão e mixagem de áudio e vídeo. Suas principais características são: ferramenta não-comercial (software livre) que disponibiliza bibliotecas de código aberto, bastante popular em sistemas Linux mas com versões compiláveis em Windows e MAC OS X, considerada veloz na conversão e gravação de arquivos de vídeo, suporta uma grande variedade de formatos de arquivos de áudio e vídeo e implementa os principais algoritmos de codificação de vídeo que permitem obter compressão dos arquivos de vídeo gerados.

- **QuickTime 7 Pro:** ferramenta comercial da Apple de captura, reprodução, gravação, conversão e mixagem de áudio e vídeo. Suportada nos sistemas operacionais Windows e MAC OS

X. Possui interface gráfica que auxilia as operações de mixagem entre áudio e vídeo. Suporta grande variedade de formatos de arquivos de áudio e vídeo e também implementa os principais algoritmos de codificação de vídeo, permitindo obter compressão dos arquivos de vídeo gerados.

5.3.1 Biblioteca de Visão Computacional "OpenCV"

O processamento digital de imagens é uma das partes fundamentais na implementação de sistemas de animação facial 2D. A qualidade visual dos quadros da animação final e a velocidade de síntese da animação são dois aspectos fundamentalmente afetados pela maneira como estes algoritmos de processamento são implementados.

Visando a implementação de um sistema piloto, este trabalho adotou a utilização da biblioteca "OpenCV" como ferramenta para implementação dos algoritmos de síntese dos quadros da animação.

A biblioteca "OpenCV" (*Open Source Computer Vision Library*) é desenvolvida em C e C++ e contém uma coleção de rotinas especializadas no processamento digital de imagens e operações de visão computacional. Esta biblioteca é suportada em ambientes Linux, Windows e MAC OS X, e possui interfaces para Python, Ruby, Matlab e outras linguagens. Atualmente esta biblioteca é livre e possui seu código aberto, contando com vasta documentação, características que possibilitam que a implementação piloto seja mais facilmente evoluída e adaptada a diferentes plataformas de execução. Originalmente, a biblioteca "OpenCV" foi desenvolvida pela Intel, com forte foco na eficiência computacional e aplicações de tempo-real. Em particular, esta biblioteca possui funções compatíveis com instruções primitivas de processadores deste fabricante, possibilitando a utilização de rotinas de "baixo-nível" otimizadas para estes processadores.

Dentre as funções de interesse desta biblioteca na implementação do sistema piloto de animação facial 2D, podem-se destacar:

- funções de manipulação de imagens (alocação em memória, cópia, criação e conversão de representação em memória dos *pixels* das imagens);
- funções de leitura e escrita de arquivos de imagens (incluindo suporte a diversos formatos de imagens e disponibilizando funções de conversão);
- rotinas de álgebra linear e manipulação de vetores e matrizes;
- rotinas básicas de processamento de imagens (amostragem e interpolação, transformações geométricas espaciais, composição de imagens);
- funções básicas de interface com o usuário (mostrar imagens na tela e manipular eventos do *mouse* e teclado).

5.4 Aspectos de Performance do Sistema Piloto

Buscando-se avaliar o sistema piloto, a Tabela 5.3 apresenta alguns dados de performance referentes à execução do aplicativo desenvolvido durante a síntese de um conteúdo pré-estabelecido.

A avaliação foi realizada fornecendo-se ao sistema de síntese, a transcrição fonética temporizada do áudio sintetizado pelo "CPqD Texto Fala", correspondente à locução da frase: "Não pode haver tréguas na guerra contra a malária".

O arquivo de áudio no formato WAV gerado para esta frase apresentava duração de 3 segundos. Considerando-se a taxa de reprodução de 30 quadros por segundo, a saída correspondente do sistema de síntese da animação é uma sequência de 90 imagens, ou quadros da animação final.

A plataforma hardware utilizada para o teste foi um computador portátil Dell, modelo Vostro 1400, com processador Intel Core 2 Duo T7250, 2 GHz e 2 Gb de memória RAM e sistema operacional Windows Vista Home Basic.

A avaliação realizada consistiu na medição do tempo total de síntese da animação associada a esta fala, medindo-se também o tempo de síntese por quadro da animação[1].

Visando-se obter o tempo médio total de síntese da animação, o processo de síntese da frase foi repetido 150 vezes. Durante a realização do teste, o computador utilizado permaneceu dedicado a esta tarefa. Manteve-se em execução concorrente, o menor número de processos do sistema operacional essenciais para o correto funcionamento da máquina. O tempo médio total de síntese da animação é apresentado na linha "Tempo total de síntese da animação" da Tabela 5.3. O símbolo σ representa o desvio padrão das medidas obtidas.

De maneira semelhante, o tempo de síntese por quadro da animação, apresentado na linha "Tempo de síntese por quadro da animação" da Tabela 5.3, foi determinado a partir da média do tempo de síntese de um quadro, considerando-se os 90 quadros resultantes da síntese da frase escolhida.

Na coluna "Implementação Original" da Tabela 5.3 são apresentados os dados e resultados obtidos considerando-se as características originais da base de imagens, em que os visemas são imagens de 200 x 150 *pixels* de dimensão, e os quadros finais da animação possuem resolução 720 x 486 *pixels* (compatível com o padrão NTSC, vide Seção 3.3).

O gráfico da Figura 5.3 mostra, através de suas marcas horizontais, o menor valor, primeiro quartil, mediana, terceiro quartil e maior valor observados no conjunto de dados analisado para o tempo de síntese de animação obtidos para a resolução 720 x 486 *pixels*.

A partir das medidas obtidas, observa-se que o sistema piloto implementado foi capaz de processar 1 segundo de animação em aproximadamente 18 segundos de síntese. As medidas obtidas permitem concluir que, considerando-se a metodologia utilizada, mais de 90% do tempo de síntese foi gasto no processo de metamorfose entre visemas-chave, enquanto que operações como a leitura e processamento do arquivo de transcrição fonética temporizada e conversão de fones em visemas consumiram menos de 1% do tempo total de síntese da animação final.

Uma das motivações deste trabalho foi a obtenção de um processo de síntese de animação facial 2D que possa ser adaptado a plataformas com capacidade de processamento e memória limitadas.

Visando avaliar a performance do sistema para displays de tamanho reduzido, incapazes de reproduzir imagens de resolução 720 x 486 *pixels*, realizou-se um experimento em que as 34 imagens originalmente selecionadas e registradas do corpus audiovisual (vide Seção 3.5) foram reduzidas para 320 x 240 *pixels*. A resolução 320 x 240 *pixels* é considerada uma simples da variação da resolução 240 x 320 *pixels*, suportada por um grande número de aparelhos celulares de tela colorida atualmente

[1] As medições foram realizadas através de funções de medição do tempo colocadas em pontos estratégicos do código do sistema piloto. Em particular, utilizou-se a função "GetTickCount()" da biblioteca C "windows.h" que, em máquinas com sistema operacional Windows, permite a obtenção do tempo real da máquina na qual o programa está sendo executado.

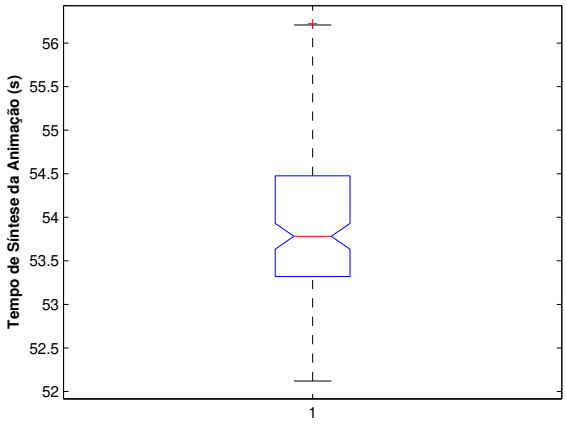


Fig. 5.3: Distribuição dos dados obtidos para o tempo total de síntese da animação, considerando-se resolução final da animação de 720 x 486 *pixels*.

Tab. 5.3: Medidas de tempo de síntese da animação utilizando sistema piloto.

	Implementação Original	Dimensões Reduzidas
Base de Imagens	34 visemas	34 visemas
Tamanho da base	3,5 Mbytes	765 quilobytes
Resolução dos visemas da base	200 x 150 *pixels*	100 x 75 *pixels*
Profundidade (bits por *pixel*)	24 bpp	24 bpp
Tempo total de síntese da animação	53,9 s (σ=0,86 s)	13,3 s (σ=0,28 s)
Tempo de síntese por quadro da animação	0,6 s (σ=0,01 s)	0,15 s (σ=0,003 s)
Resolução dos quadros da animação final	720 x 486 *pixels*	320 x 240 *pixels*

existentes. A Figura 5.5 mostra a visualização de uma animação com resolução 320 x 240 *pixels* sintetizada em um computador e reproduzida em um aparelho celular.

As operações de processamento para redução dos visemas da base de imagens foram implementadas através das ferramentas de processamento em lote disponibilizadas pelo aplicativo de processamento de imagens IrfanView. Inicialmente, definiu-se em uma das imagens (720 x 486 *pixels*) as coordenadas de uma janela de recorte de tamanho 640 x 480 *pixels*. A janela de recorte determinou uma região de interesse que possui como elemento central a face da apresentadora, possibilitando descartar as bordas externas à janela. A mesma janela de recorte foi aplicada a todas as 34 imagens. Em seguida, realizou-se a operação de redimensionamento, em que as imagens foram reduzidas à metade, utilizando-se o algoritmo de interpolação de *pixels* Lanczos.

Finalmente, as imagens reduzidas (320 x 240 *pixels*) foram processadas extraindo-se a região de interesse correspondente aos visemas a serem armazenadas na base (vide Seção 3.5.3) e medindo-se os pontos-âncora utilizados durante o processo de síntese (vide Seção 3.6). O resultado destas operações foi a construção de uma base de 34 visemas com dimensões reduzidas de 100 x 75 *pixels*. As imagens foram salvas em formato BMP, sem compressão, com profundidade de 24 bpp (bits por pixel, padrão RGB). Nesta configuração, a base de imagens ocupa aproximadamente 765 kilobytes de memória estática (vale destacar que este tamanho de base de imagens pode ser armazenada por uma grande gama de dispositivos móveis portáteis ou similares atualmente existentes, que possuem capacidade limitada de memória quando comparados a sistemas desktop).

Após a construção da nova base de imagens, o sistema piloto foi novamente avaliado durante a síntese da mesma frase anteriormente utilizada. Nesta configuração, os quadros finais da animação sintetizados pelo sistema possuem resolução 320 x 240 *pixels*.

Nesta versão, o sistema piloto foi capaz de gerar 1 segundo de animação em aproximadamente 4,4 segundos de tempo de síntese. O gráfico da Figura 5.4 mostra, através de suas marcas horizontais, o menor valor, primeiro quartil, mediana, terceiro quartil e maior valor observados no conjunto de dados analisado para o tempo de síntese de animação obtidos para a resolução 320 x 240 *pixels*.

A coluna “Dimensões Reduzidas” da Tabela 5.3 apresenta os resultados de performance medidos a partir da base de visemas de dimensões reduzidas. A tabela mostra que ao se reduzir as dimensões das imagens da base de visemas em 50%, obteve-se uma redução no tempo total de síntese da animação de aproximadamente 75%. É possível correlacionar a redução no tempo de síntese à diferença entre o número total de *pixels* processados durante a etapa de metamorfose entre visemas nas duas configurações da base de imagens. Na base de visemas original, em que os visemas armazenados possuem dimensões 200 x 150 *pixels*, tem-se um total de 30.000 *pixels* processados. Na versão com dimensões reduzidas (100 x 75 *pixels*) tem-se 7.500 *pixels* processados, ou 25% do número total de *pixels* das imagens da base de visemas original.

5.5 Comentários Finais

Este capítulo apresentou alguns dos aspectos mais importantes da implementação de um sistema piloto de animação facial 2D, seguindo o processo de síntese proposto neste trabalho.

A partir do sistema piloto foi possível gerar, de maneira automatizada, as animações utilizadas durante a avaliação do processo de síntese (abordada no Capítulo 6). Esta implementação piloto, além de permitir a validação do abordagem proposta, estabelece também uma ferramenta útil de

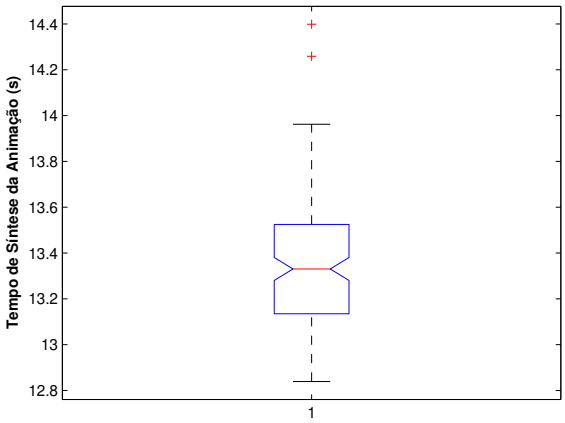


Fig. 5.4: Distribuição dos dados obtidos para o tempo total de síntese da animação, considerando-se resolução final da animação de 320 x 240 *pixels*.

Fig. 5.5: Animação facial 2D apresentada em aparelho celular (Nokia N95) – resolução 320 x 240 *pixels*

análise e suporte à evolução do sistema, uma vez que novos algoritmos e funcionalidades podem ser facilmente adaptados e testados a partir da plataforma base implementada neste trabalho.

O sistema piloto, tal como implementado, não teve como foco principal o desenvolvimento de uma aplicação de animação de alta performance e/ou síntese em tempo real. Buscou-se, inicialmente, o sucesso na implementação dos algoritmos correspondentes aos passos de síntese da animação apresentados no Capítulo 4.

As medidas de tempo de síntese apresentadas na Seção 5.4 permitiram realizar uma avaliação preliminar da implementação piloto apresentada, sob o ponto de vista da síntese de frases curtas como a utilizada durante a realização dos testes. Neste contexto, é possível afirmar que o sistema piloto implementado apresenta um tempo de latência de síntese de animação que pode ser acomodado satisfatoriamente a diversas aplicações que não demandem a síntese em tempo real. Estas aplicações são caracterizadas, por exemplo, por um baixo nível de interatividade, em que o tempo de processamento e síntese da animação seja transparente para o usuário final. Exemplos de aplicações deste tipo são: animações para treinamento ou tutoriais, conteúdo de notícias ou propaganda disponibilizados em repositórios e entregues sob demanda ao usuário, aplicações web em que o tempo de latência de síntese da animação se confunda com outras operações relacionadas à aplicação.

Em muitas aplicações, porém, a síntese de uma *talking head* deve ser realizada em tempo real. Este é o caso, por exemplo, de aplicações que envolvam uma situação de diálogo em que, uma vez provocado um estímulo, como uma pergunta, o agente virtual deve responder em um intervalo de tempo suficientemente curto para que seu comportamento seja considerado próximo ao de um interlocutor real.

É válido ressaltar, no entanto, que o tamanho das frases sintetizadas, bem como o número de visemas a serem selecionados da base de imagens, são fatores que impactam o tempo de síntese da animação final.

Uma avaliação mais abrangente da performance computacional da implementação piloto se faz necessária para o levantamento dos principais pontos de otimização que possibilitem a implementação de um sistema de síntese de animação facial em tempo real. Dentre as possíveis iniciativas de redução no tempo de síntese da animação a serem aplicadas na implementação piloto, podem-se citar: a implementação de algoritmos de solução de sistemas lineares otimizados para a aplicação considerada, a otimização de funções de processamento de imagens através da utilização funções primitivas de recursos de hardware especializados em processamento gráfico, a redução do processamento necessário para a metamorfose entre visemas-chave através da redução da resolução e/ou profundidade (número de camadas e bits utilizados para representar cada *pixel* de uma imagem) das imagens da base. Tais otimizações são consideradas propostas para trabalhos futuros relacionadas à evolução do sistema.

Capítulo 6

Teste de Inteligibilidade da Fala

6.1 Introdução

Avaliar a animação gerada por um sistema de animação facial é um dos passos essenciais de seu processo de desenvolvimento, permitindo o estabelecimento de parâmetros de comparação e o acompanhamento do progresso obtido com novas versões ou abordagens alternativas que venham a ser exploradas. Os sistemas existentes empregam diferentes técnicas na avaliação de seus resultados, não existindo critérios universalmente aceitos para estabelecer a qualidade de uma animação. A definição do tipo de avaliação a ser realizada é tipicamente baseada nos objetivos do teste e as características de cada sistema.

Uma das abordagens de avaliação é a realização de testes subjetivos onde observadores expressam suas opiniões a respeito de determinados aspectos da animação.

Em (PANDZIC; OSTERMANN; MILLEN, 1999), por exemplo, implementa-se um serviço interativo em que uma face animada é utilizada para fornecer informações aos usuários. Após a utilização do serviço, os usuários são convidados a expressar seu grau de satisfação em relação ao mesmo. Verifica-se neste trabalho o foco na investigação da contribuição da face animada na usabilidade de serviços interativos. Os resultados obtidos são comparados com interfaces puramente textuais ou que disponibilizam o vídeo real de uma face.

No sistema apresentado em (COSATTO; GRAF; OSTERMANN, 2004), aspectos como "naturalidade" da face falante, "sincronismo", "suavidade" e "precisão" da fala na animação são avaliados solicitando-se que os observadores forneçam sua opinião através de uma escala de opinião de 0 a 5, do tipo MOS (*Mean Opinion Score*). Este tipo de teste subjetivo tem suas origens na avaliação da qualidade de voz para telefonia (ITU, 1996). No caso deste sistema, tem-se como objetivo avaliar, entre outros aspectos, a reprodução de sinais de comunicação não-verbais, ou prosódia visual.

Embora a realização de testes subjetivos forneçam informações importantes sobre a qualidade de uma animação facial, as respostas apresentadas podem apresentar grande variabilidade. A face humana, em particular, pode despertar reações variadas de acordo com sua aparência e as feições apresentadas. Um observador pode julgar uma face "arrogante" ou "simpática", ou o estilo do cabelo pode transmitir idéias que influenciem sua opinião. Por este motivo, testes subjetivos tornam-se custosos, devendo ser realizados com um grande número de observadores, de variados grupos demográficos e com um grande número de amostras.

Uma alternativa que tem sido empregada em sistemas de animação facial 2D é a realização do

teste de Turing (TURING, 1950), no qual os observadores são simplesmente convidados a distinguir visualmente faces reais de faces animadas em trechos de vídeo com a mesma locução (EZZAT; GEIGER; POGGIO, 2002), (BRAND, 1999). Este tipo de teste procura fornecer uma indicação do nível de vídeo-realismo alcançado por uma animação, entretanto, ele não é capaz de fornecer informações adicionais que apontem possíveis pontos de melhoria.

Uma outra alternativa, particularmente interessante no contexto deste trabalho, são os testes objetivos de inteligibilidade da fala que têm sido empregados, com pequenas variações, para a avaliação de sistemas de animação facial sincronizada à fala (BENOÎT; GOFF, 1998), (PANDZIC; OSTERMANN; MILLEN, 1999), (GEIGER; EZZAT; POGGIO, 2003), (DE MARTINO; VIOLARO, 2007). O objetivo deste tipo de teste, derivado da proposta de Sumby e Pollack (1954), é avaliar a contribuição visual da imagem à inteligibilidade da fala sob diferentes condições de degradação do áudio. Esta avaliação fornece informações diretas sobre a qualidade de reprodução dos movimentos articulatórios visíveis da fala, incluindo a sincronia e harmonia com o áudio. Adicionalmente, a obtenção de resultados objetivos facilita a comparação entre diferentes versões ou abordagens empregadas durante o processo de evolução do sistema.

Visando avaliar os resultados obtidos pelo sistema apresentado neste trabalho, foram realizados testes de inteligibilidade da fala, descritos neste capítulo. A escolha por este tipo de abordagem reflete o foco do trabalho na reprodução vídeo-realista dos movimentos articulatórios visíveis da fala, e em particular, da reprodução dos efeitos de coarticulação através de visemas dependentes de contexto, que afetam diretamente a contribuição visual da animação à inteligibilidade da fala.

O teste realizado utilizou como objeto de avaliação um conjunto de 27 palavras sem significado, ou logatomas, que, em ordem aleatória, foram apresentados em três versões diferentes a cada observador participante do teste. A primeira versão é composta apenas do áudio da locução dos logatomas. A segunda versão é constituída do vídeo, e áudio associado, de um locutor real produzindo os logatomas. Já a terceira versão é composta pela animação 2D sintetizada e sincronizada com a fala. Para cada uma das três versões, o áudio associado foi contaminado com ruído branco, produzindo três níveis de relação sinal-ruído (SNR - *Signal to Noise Ratio*) distintos: -12 dB, -18 dB e -24 dB. Os participantes, após observarem/ouvirem uma apresentação, tiveram por tarefa indicar o logatoma produzido através de seleção numa lista de opções. Os resultados do teste são expressos pela análise estatística da quantidade de respostas corretas fornecidas pelos participantes para cada versão de apresentação e nível de degradação de ruído. Os resultados permitem avaliar a contribuição da animação facial à inteligilidade da fala, e assim a qualidade de reprodução da movimentação articulatória visível, em relação a situação em que não há informação visual (apresentação apenas do áudio) e o caso ideal (apresentação do vídeo de um locutor real).

Nas seções 6.2, 6.3 e 6.4 deste capítulo, são descritos os detalhes do material de avaliação, o protocolo de teste utilizado e o perfil dos participantes do teste. Em seguida, as seções 6.5 e 6.6 apresentam os resultados obtidos e as análises decorrentes destes resultados. Finalmente, o Apêndice C fornece informações teóricas sobre o tratamento estatístico utilizado na análise dos dados.

6.2 Preparação do Material de Teste

A avaliação de inteligibilidade da fala teve como material base 27 palavras sem significado, ou logatomas, encapsuladas em uma frase veículo com a seguinte estrutura: “Ela fala <*logatoma*>”. A

frase veículo tem como função preparar o observador para o momento de pronúncia do logatoma.

Foram utilizados no teste de inteligibilidade logatomas paroxítonos do tipo 'CVCV, com C=/p, t, k, s, ʃ, l, λ, ɣ/, V=/i,a,u/. Os logatomas formados pela concatenação de dois contextos CV foram escolhidos para estimular a produção dos efeitos da coarticulação durante a pronúncia.

O conjunto de consoantes foi selecionado elegendo-se um representante de cada agrupamento homofema da Tabela 3.3.

O áudio original da fala da apresentadora, correspondente à pronúncia do logatoma, foi extraído do corpus audiovisual. Os 27 arquivos de áudio resultantes foram processados utilizando-se a ferramenta Adobe Audition seguindo-se os passos:

- sub-amostragem dos arquivos de áudio originalmente capturados a 44 kHz para 16 kHz;
- formatação do arquivo de áudio visando a uniformização da duração do áudio correspondente aos diversos logatomas;
- operação de restauração de clip (função disponibilizada pela ferramenta Adobe Audition™);
- concatenação com o trecho de áudio correspondente à frase veículo "Ela fala", também pronunciada por voz feminina;
- contaminação com ruído produzindo 3 diferentes condições de relação sinal-ruído: -12 dB, -18 dB e -24 dB.

Para o processo de adição de ruído aos arquivos de áudio, os mesmos foram convertidos para o formato *float* e normalizados no intervalo [-1,0;1,0] após uma divisão por 32768. Um sinal de ruído de distribuição uniforme (ruído branco) foi adicionado às frases veículo. O nível de SNR foi calculado considerando-se somente o trecho de áudio correspondente à locução do logatoma. Após a adição do ruído, a amplitude do sinal foi novamente normalizada para o intervalo [-1,0;1,0] e quantizada novamente para 16 bits de resolução. O sinal resultante foi salvo em arquivos do tipo RIFF (*Resource Interchange File Format*) WAVE.

Após o processamento dos áudios originais dos logatomas, teve-se como resultado 81 arquivos de áudio, correspondentes a 27 arquivos em 3 diferentes versões de degradação do áudio.

Em seguida, as cópias de áudio degradadas foram ressincronizadas ao vídeo original da apresentadora. Da maneira que o vídeo foi editado, os lábios da apresentadora permanecem em repouso durante o início da frase veículo e se movimentam somente durante a pronúncia do logatoma, em sincronia com o áudio.

A partir do material de transcrição fonética originalmente gerado pela segmentação manual do material do corpus audiovisual, geraram-se animações correspondentes à pronúncia dos logatomas. De maneira similar à realizada para o vídeo gravado da face real, realizou-se a operação de sincronia entre os trechos de áudio degradado e os trechos de animação.

Para edição dos arquivos de vídeo utilizou-se a ferramenta Quicktime Pro, e o material resultante foi encapsulado em arquivos de extensão ".mov" sem nenhuma compressão de áudio e compressão de vídeo H.624, no padrão NTSC.

Concluiu-se assim a preparação do material utilizado na avaliação de inteligibilidade, resultando na geração de 243 arquivos organizados em 3 grupos:

- **Somente Áudio:** 81 arquivos de aúdio correspondentes em 3 versões de degradação do áudio para cada um dos 27 logatomas;
- **Animação+Áudio:** 81 arquivos de vídeo correspondentes a animações faciais em 3 versões de degradação do áudio para cada um dos 27 logatomas
- **Vídeo+Áudio:** 81 arquivos de vídeo em 3 versões de degradação do áudio para cada um dos 27 logatomas.

Os 243 arquivos resultantes foram organizados em 9 grupos, cada um contendo 27 frases com a mesma qualidade acústica (mesmo nível de SNR) e de mesma natureza (áudio somente, animação+áudio ou vídeo+áudio). Durante a apresentação do material aos observadores, os grupos eram apresentados de maneira aleatória, bem como cada um dos 27 logatomas pertencentes a cada grupo.

6.3 Protocolo de Teste

O teste de inteligibilidade foi conduzido em uma sala com baixo nível de ruído e isolamento acústico. Uma ferramenta software foi desenvolvida para a apresentação, coleção e registro dos votos dos observadores sujeitos ao teste. A aplicação Java foi executada em plataforma Dell Optiplex GX270, com processador Intel Pentium 4 (3 GHz). A apresentação do material foi visualizada através de um monitor de 17 polegadas e o áudio era reproduzido através de fones de ouvido de alta qualidade. A Figura 6.3 mostra a tela utilizada para apresentação do material de teste.

Fig. 6.1: Tela da ferramenta utilizada para apresentação e votação do teste de inteligibilidade da fala

Após a apresentação de uma frase veículo, o observador era solicitado a indicar o logatoma compreendido selecionando uma das 28 opções disponíveis, composta de 27 logatomas e uma opção NDA (Nenhuma das Anteriores). A Figura 6.2 mostra o painel de votação em detalhes. Após a seleção da opção, o participante deveria confirmar seu voto e prosseguir para a próxima apresentação.

Os observadores foram encorajados a escolher uma opção mesmo quando estivessem em dúvida entre diferentes opções de logatomas e optar pela alternativa NDA somente quando eles de fato não tivessem nenhuma suspeita de qual logatoma foi apresentado, ou quando julgassem que o que foi escutado não correspondia a nenhuma das opções de logatoma apresentadas. Os participantes foram informados de que os logatomas eram do tipo 'CVCV e que logatomas diferentes dos mostrados nas opções de votação poderiam ser apresentados.

O tempo médio de realização do teste foi de 30 minutos.

Fig. 6.2: Detalhe do painel de votação

6.4 Características da População de Participantes

O teste de inteligibilidade da fala contou com a participação de 41 pessoas, sendo 14 participantes do sexo feminino e 27 participantes do sexo masculino. Os participantes, todos funcionários da Fundação CPqD e sem envolvimento direto com o projeto de animação facial, declararam condições normais de visão e audição. Os participantes tinham idade variando de 18 a 54 anos, com idade média de 33 anos. As análises dos dados não levaram em conta o genêro e a idade dos participantes.

6.5 Resultados

A Tabela 6.1 apresenta o resumo da compilação das respostas fornecidas pelos participantes. Na tabela, a coluna "Média" indica a média da quantidade de acertos normalizada entre 1 (100% de acerto) e 0 (nenhum acerto).

Os valores médios de acertos em termos percentuais são também apresentados no gráfico da Figura 6.3. A partir deste gráfico, é possível visualizar que a porcentagem de acertos para os conteúdos "Somente Áudio", "Animação + Áudio" e "Vídeo + Áudio" é tanto maior quanto menor o nível de degradação do áudio, expresso por maiores valores de SNR. Em outras palavras, a maior possibilidade de compreensão do áudio residual, em meio ao ruído, reflete-se diretamente na maior taxa de acertos nos três tipos de conteúdo apresentados. De fato, é possível perceber no gráfico a tendência das taxas de acertos para os três tipos de conteúdo serem cada vez mais semelhantes, conforme diminui-se a presença do ruído no áudio. Esta tendência é compatível com os resultados dos trabalhos (SUMBY; POLLACK, 1954) e (DE MARTINO; VIOLARO, 2007), que incluíram testes com níveis de SNR adicionais.

Do gráfico da Figura 6.3, é possível também avaliar a contribuição na inteligibilidade da fala das animações geradas pelo sistema implementado. Considerando-se a situação de maior degradação do áudio, com SNR de -24 dB, observa-se uma baixíssima porcentagem de acertos para o conteúdo "Somente Áudio". De fato, tal nível de degradação do áudio pode ser comparado à situação de ausência total de áudio, na qual, na prática, a inteligibilidade da fala é conseguida somente a partir de um exercício de leitura orofacial. Neste caso, observa-se que o ganho na inteligibilidade da fala promovido pela animação foi de 24%, enquanto para o vídeo real foi de 38%. Desta maneira, observa-se a efetiva contribuição da informação visual da animação à inteligibilidade da fala, em que a solução apresentada neste trabalho encontra-se a aproximadamente 63% da contribuição à inteligibilidade da fala promovida pelo vídeo de uma face real.

Para os níveis de SNR de -18 dB e -12 dB, a principal característica observável é a proximidade entre as porcentagens de acertos para os conteúdos "Animação + Áudio" e "Vídeo + Áudio". Considerando-se a situação do áudio com SNR de -18 dB, observa-se que a animação facial fornece um ganho na inteligibilidade da fala de aproximadamente 33% em relação à situação em que apenas o áudio é apresentado. O ganho fornecido pelo vídeo real, por sua vez, foi de 40%. Para SNR de -12 dB os ganhos são de aproximadamente 25% e 30% para a animação e o vídeo real, respectivamente.

Tab. 6.1: Resultados do Teste de Inteligibilidade

SNR	**Somente Áudio**		**Animação + Áudio**		**Vídeo + Áudio**	
	Média	Variância	Média	Variância	Média	Variância
-12 dB	0,3424	0,0077	0,5944	0,0136	0,6396	0,0145
-18 dB	0,1572	0,0069	0,4860	0,0191	0,5601	0,0128
-24 dB	0,0072	0,0004	0,2439	0,0132	0,3839	0,0136

Análises adicionais podem ser derivadas da análise de variância (ANalysis Of Variance - ANOVA) dos dados obtidos. Essencialmente, a análise de variância permite determinar a probabilidade p de que conjuntos independentes de amostras possuam a mesma média, sendo esta hipótese definida como hipótese nula (H_0) (veja Apêndice C). A rejeição desta hipótese permite afirmar que os conjuntos de amostras analisados podem ser considerados estatisticamente discerníveis.

Considerando-se os conjuntos de amostras obtidos para os conteúdos "Somente Áudio", "Anima-

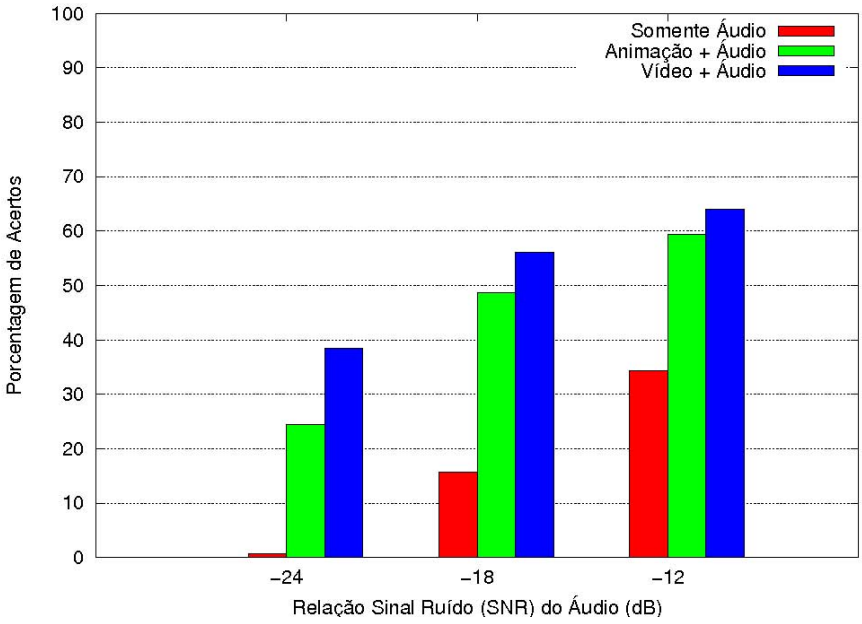


Fig. 6.3: Porcentagem de Acertos em função do nível de SNR

Tab. 6.2: Resultados da Análise de Variância

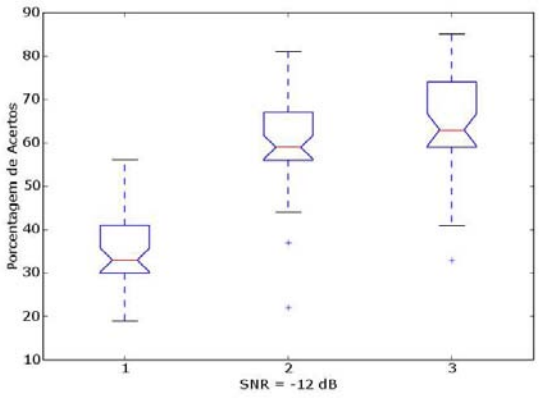
SNR	**Hipótese nula** H_0	**F**	**p**
-12 dB	$\mu_a = \mu_{a+a} = \mu_{v+a}$	88,3376	$< 0,001$
	$\mu_a = \mu_{a+a}$	122,8871	$< 0,001$
	$\mu_a = \mu_{v+a}$	163,6644	$< 0,001$
	$\mu_{a+a} = \mu_{v+a}$	2,8801	$0,0936$
-18 dB	$\mu_a = \mu_{a+a} = \mu_{v+a}$	144,5427	$< 0,001$
	$\mu_a = \mu_{a+a}$	169,7029	$< 0,001$
	$\mu_a = \mu_{v+a}$	332,3711	$< 0,001$
	$\mu_{a+a} = \mu_{v+a}$	6,8227	$0,0107$
-24 dB	$\mu_a = \mu_{a+a} = \mu_{v+a}$	163,9649	$< 0,001$
	$\mu_a = \mu_{a+a}$	171,4500	$< 0,001$
	$\mu_a = \mu_{v+a}$	413,1182	$< 0,001$
	$\mu_{a+a} = \mu_{v+a}$	29,5926	$< 0,001$

ção + Áudio" e "Vídeo + Animação", define-se:

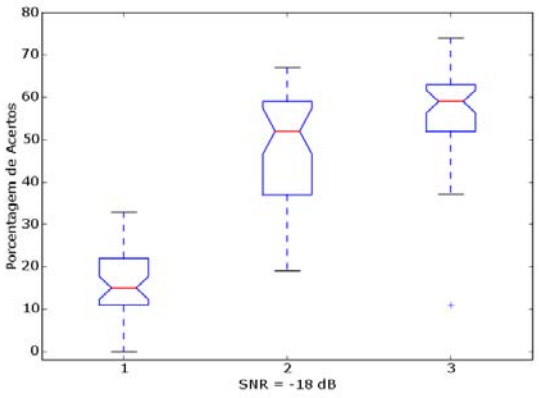
- μ_a: média da taxa de acertos para "Somente Áudio";
- μ_{a+a}: média da taxa de acertos para "Animação + Áudio";
- μ_{v+a}: média da taxa de acertos para "Vídeo + Áudio".

A Tabela 6.2 apresenta os resultados numéricos obtidos para a análise de variância conduzida para diferentes hipóteses nulas [1].

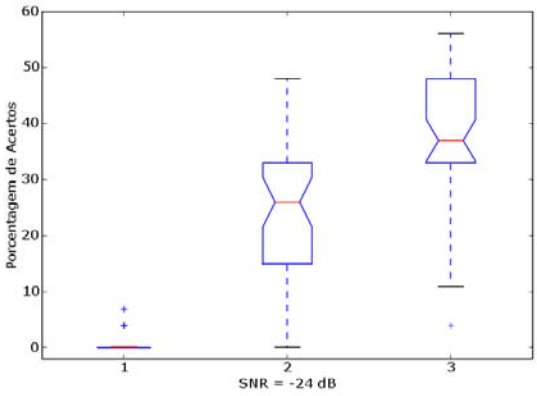
Adicionalmente, os gráficos da Figura 6.4 auxiliam a interpretação destes resultados e essencialmente mostram, através de suas marcas horizontais, o menor valor, primeiro quartil, mediana, terceiro quartil e maior valor observados no conjunto de dados analisado.



(a) SNR = -12 dB



(b) SNR = -18 dB



(c) SNR = -24 dB

Fig. 6.4: Distribuição dos dados para SNR = -12 dB, -18 dB e -24 dB: (1) Somente Áudio; (2) Animação + Áudio; (3) Vídeo + Áudio

[1] Estes resultados foram obtidos através da função *anova1()* da ferramenta Matlab™, que utiliza um intervalo de 95% de confiança para teste da hipótese nula.

Na Tabela 6.2, a primeira hipótese nula testada para os 3 níveis de SNR avaliados é $H_0 : \mu_a = \mu_{a+a} = \mu_{v+a}$, resultando em valores de p<0,001 e indicando que os conjuntos de amostras obtidos não são provenientes de um mesmo fenômeno com distribuição normal. Em seguida, pares de conjuntos de amostras foram analisados entre si.

A partir dos gráficos das figuras 6.4(a) e 6.4(b), é possível observar que as medianas das amostras "Animação + Áudio" e "Vídeo + Áudio" para os níveis de SNR de -12 dB e -18 dB se mostram próximas. De fato, o teste da hipótese $H_0 : \mu_{a+a} = \mu_{v+a}$ para estes níveis demonstra que estes pares de amostras não podem ser considerados estatiscamente discerníveis e que os dados podem ser interpretados estatiscamente como flutuações casuais de uma mesma distribuição normal.

A mesma situação, no entanto, não é observada para o nível de degradação do áudio de -24 dB. O gráfico da Figura 6.4(c) mostra o maior distanciamento nas distribuições dos resultados de animação e vídeo e a análise de variância para todos os pares de amostras resulta em significância estatística p<0,001 (Tabela 6.2).

Tais resultados permitem definir um importante ponto referencial para evolução futura da solução apresentada neste trabalho, mostrando, de maneira objetiva, o impacto causado no nível de inteligibilidade da fala pelas aproximações realizadas pelo processo de síntese de animação a partir de uma base de imagens reduzida. Nas situações de SNR de -18 dB e -12 dB, animação e vídeo real apresentam contribuição semelhante, de tal maneira que, estatiscamente, não é possível realizar distinção entre os dois conteúdos. Já para a situação de SNR = -24 dB, em que a informação visual é essencial para a inteligilibidade dos conteúdos apresentados, o sistema implementado está a aproximadamente 63% da situação ideal, caracterizada pelo vídeo de uma face real.

6.6 Comentários Finais

A partir dos resultados do teste de inteligibilidade da fala, é possível afirmar que:

- A informação da movimentação articulatória visível apresentada pela animação facial contribui para a inteligilidade da fala, principalmente em situações de forte contaminação do áudio por ruído.
- A abordagem adotada reproduz características relevantes da movimentação articulatória visível em consonância com a fala.
- Para situações de contaminação do áudio por ruído em níveis de SNR maiores que -18 dB, a solução apresentada é capaz de gerar animações vídeo-realistas no que se refere à contribuição à inteligibilidade da fala, a partir de uma base de apenas 34 visemas dependentes de contexto.
- Os resultados obtidos para o nível SNR = -24 dB indicam que a solução apresentada está a 63% da contribuição à inteligibilidade da fala promovida por um vídeo real, definindo o espaço de aperfeiçoamento do sistema no auxílio à capacidade de leitura orofacial. Os resultados obtidos são importantes referenciais para a avaliação de futuras versões do sistema.

A versão do sistema utilizada para geração das animações do teste conduzido sintetiza animações a partir de uma reduzida base de imagens. Este tipo de versão é particularmente indicada para utilização em dispositivos com baixa capacidade de processamento e memória. Tais dispositivos tipicamente

apresentam telas de tamanho reduzido e restrição na resolução de imagens mostradas, ambos fatores que influenciam a qualidade da visualização e, portanto, as características da contribuição visual das imagens. Assim sendo, sugere-se como trabalho futuro a realização de testes de inteligibilidade da fala com a apresentação de conteúdos a partir de dispositivos deste tipo. Um dos objetivos desta modalidade de avaliação seria avaliar o compromisso entre o atual nível de qualidade da animação versus custos adicionais associados à evolução do sistema, tais como maiores custos computacionais ou de armazenamento de imagens.

Por último, vale ressaltar que o tipo de teste conduzido neste trabalho não leva em consideração aspectos importantes relacionados à "naturalidade" da face falante, tais como a reprodução de gestos e sinais de comunicação não-verbais ou movimentos fisiológicos da face. Avaliações que envolvam estes aspectos devem ser consideradas no processo de evolução e avaliação futura do sistema.

Capítulo 7

Conclusões

Agentes virtuais e *talking heads* encontram aplicações em diversas áreas, tais como: entretenimento, comunicação pessoal, sistemas de auxílio à navegação, apresentação de notícias, comércio eletrônico, educação e treinamento, e interfaces humano-computador. Juntamente com as áreas de reconhecimento e síntese de fala e inteligência artificial, a pesquisa na área de animação facial é essencial para a obtenção de personagens virtuais cada vez mais realistas, cuja implementação seja viável nas mais diversas plataformas, como computadores pessoais, dispositivos móveis portáteis, eletrodomésticos e máquinas de atendimento automático, entre outras.

O sistema de animação facial apresentado neste trabalho representa um passo na direção da obtenção de animações faciais vídeo-realistas. A solução projetada e implementada permite a geração de animação facial em sincronia e harmonia com a fala proveniente de processo de gravação ou de síntese por computador da voz humana, caracterizando uma *talking head*. Em particular, a língua considerada no contexto deste trabalho é o Português do Brasil.

Os focos principais deste trabalho foram a reprodução da movimentação articulatória visível da fala e a proposta de uma metodologia de síntese capaz de ser adaptada a sistemas com recursos computacionais de processamento e memória limitados.

O processo de síntese desenvolvido adota a abordagem de animação facial baseada em imagens, ou 2D, em que a modelagem da face e cabeça é implementada através de imagens fotográficas capturadas de uma face real. A partir desta abordagem, este trabalho dedicou atenção especial à reprodução dos efeitos da coarticulação presentes na movimentação articulatória visível da fala. A coarticulação é a alteração observada no padrão articulatório de um determinado segmento da língua resultante da influência da pronúncia de segmentos adjacentes ou próximos.

Para isso, o trabalho destacou como etapa essencial da implementação do sistema o processo de construção de uma base de imagens de visemas, definidas como imagens estáticas, visualmente contrastantes entre si, representando as configurações típicas de articulação dos diversos segmentos da fala. Para esta implementação, o sistema baseou-se no trabalho pioneiro de DE MARTINO (2005), que identificou visemas dependentes de contexto para o Português do Brasil e aplicou esta técnica a um sistema de animação facial baseado em modelo.

No contexto deste trabalho, visemas dependentes de contexto são definidos como imagens de posturas labiais estáticas associadas não somente à produção de um segmento isolado, mas influenciadas pela interferência de segmentos específicos que o antecedem e sucedem em uma sequência de locução. Em outras palavras, as imagens utilizadas neste trabalho, associadas aos visemas dependen-

tes de contexto, foram capturas durante a locução de segmentos em contextos fonéticos previamente definidos.

Adicionalmente, visando uma implementação de baixo custo computacional e adaptável a diferentes plataformas, o processo de síntese foi implementado utilizando-se a técnica de metamorfose entre visemas, popular em sistemas de animação facial 2D (Seção 2.4.1). Esta abordagem é caracterizada principalmente por uma base de imagens de tamanho reduzido e a aplicação de modelos da movimentação articulatória simplificados.

A adoção de visemas dependentes de contexto combinada à técnica de metamorfose entre visemas caracteriza uma estratégia alternativa e inovadora de implementação de um sistema de animação facial 2D, capaz de contemplar os efeitos da coarticulação a partir de uma base de imagens reduzida de apenas 34 imagens.

A base de imagens implementada neste trabalho foi construída a partir da captura, em condições controladas, de um corpus audiovisual de uma face feminina pronunciando conteúdos previamente definidos. Como parte essencial da técnica proposta, o trabalho descreveu um processo de análise, seleção, pré-processamento e rotulação manual de 34 visemas dependentes de contexto extraídas do universo de imagens capturadas pelo corpus audiovisual.

A síntese da animação é implementada tendo-se como parâmetro de entrada a transcrição fonética temporizada da fala a ser visualmente animada. Na implementação piloto apresentada neste trabalho, a fala é gerada por um sistema de conversão texto-fala, que também fornece a transcrição fonética temporizada. Em seguida, a sequência de fonemas da transcrição é convertida em uma sequência de visemas dependentes de contexto, caracterizando a aplicação do modelo de coarticulação. Os visemas convertidos representam poses-chave, ou visemas-chave, da animação final. Os visemas-chave são processados através de um algoritmo de metamorfose entre imagens, utilizando-se apenas 5 pontos de controle para guiar o processo de *warping*, implementado utilizando-se funções de base radial.

O processo de metamorfose entre visemas é guiado temporalmente por uma função não linear de transição, que procura acomodar a dinâmica da movimentação articulatória visível. Esta abordagem caracteriza uma estratégia alternativa à transição linear tipicamente adotada por sistemas baseados na metamorfose entre visemas. Em particular, diferentemente da implementação de Edge e Maddock (2003), que também utiliza funções não lineares de transição, o sistema aqui apresentado utiliza um menor número de visemas na sua base de imagens e um menor número de pontos-âncora para o processo de metamorfose, e adota curvas cúbicas paramétricas de Hermite, numa abordagem mais direta e computacionalmente mais eficiente, já que não são necessários processos de análise de contextos fonéticos como pré-requisito para determinar parâmetros da função de transição (Seção 4.4).

Este trabalho inclui também a implementação de um sistema piloto de síntese da animação facial 2D, que permitiu avaliar a viabilidade e a qualidade visual do processo de síntese apresentado.

A partir da análise do o tamanho da base de imagens e tempo de síntese apresentados pelo sistema piloto (Capítulo 5), foi possível concluir que o sistema pode ser adaptado a plataformas de capacidade de processamento e memória limitados, como telefones celulares ou PDAs (*Personal Digital Assistants*), em aplicações que não exijam a síntese em tempo real.

As animações geradas pelo sistema piloto foram utilizadas para a realização de testes de inteligibilidade da fala, em que se buscou avaliar a contribuição da animação facial para o aumento da inteligibilidade da fala em condições de áudio desfavoráveis, fortemente degradado por ruído. Os resultados apresentados no Capítulo 6 mostram que as animações geradas pelo sistema favorecem o aumento da inteligibilidade da fala nestas condições. Em particular, foi possível estabelecer que,

para situações em que o áudio se encontra fortemente degradado por ruído, as animações geradas a partir de apenas 34 visemas dependentes de contexto encontram-se a aproximadamente 63% da contribuição à inteligibilidade da fala promovida pelo vídeo de uma face real.

A partir das características do sistema apresentado, é possível revisitar o gráfico da Figura 2.10 apresentado no Capítulo 2, e situar este trabalho no universo de sistemas de animação facial 2D existentes segundo os critérios de flexibilidade e realismo, como mostra o gráfico da Figura 7.1.

A partir deste gráfico, é possível destacar que o presente trabalho representa um passo na direção de sistemas de animação facial 2D que apresentam alto nível de realismo e alto nível de flexibilidade. Quando comparado a sistemas como (SCOTT et al., 1994) e (EZZAT; POGGIO, 1998), o sistema apresentado é capaz de fornecer maiores níveis de realismo a partir da reprodução dos efeitos da coarticulação e da modelagem não linear aplicada à transição entre visemas-chave durante o processo de metamorfose. Por outro lado, o sistema apresenta maior flexibilidade que sistemas como *Video Rewrite* (BREGLER; COVELL; SLANEY, 1997), (COSATTO; GRAF, 2000), (EZZAT; GEIGER; POGGIO, 2002) e (EDGE; MADDOCK, 2003), pois implementa estratégias de síntese menos complexas e base de imagens de tamanho reduzido.

Fig. 7.1: Comparação entre este trabalho e outros sistemas de animação facial 2D segundo critérios de flexibilidade e realismo

Considerando-se os aspectos apresentados, as principais contribuições deste trabalho são:

- o desenvolvimento de um processo de síntese de animação facial 2D baseado na metamorfose entre visemas dependentes de contexto, capaz de contemplar os efeitos da coarticulação a partir de uma base de imagens reduzida, com apenas 34 imagens;
- a implementação de um sistema de animação facial 2D para o Português do Brasil (vale ressaltar que a autora desconhece trabalhos anteriores que tratem da implementação de sistemas desta natureza para esta língua);

- complementar a investigação realizada por DE MARTINO (2005) ao utilizar visemas dependentes de contexto para reprodução da movimentação articulatória visível da fala em um sistema 2D;
- a apresentação de uma estratégia alternativa de transição não linear entre visemas em um sistema de animação facial 2D baseado na metamorfose entre visemas;
- a adoção de uma abordagem de síntese da animação de baixo custo computacional, que pode ser adaptada a dispositivos com capacidade limitada de processamento e armazenamento de dados;
- a obtenção de um sistema de síntese de animação facial capaz de sintetizar animações com nível de vídeo-realismo que favorece a inteligibilidade da fala em condições desfavoráveis de áudio.

Partindo-se da solução apresentada e levando-se em conta suas limitações, várias são as frentes de trabalho que impulsionam a evolução do sistema de síntese apresentado visando a obtenção de animações faciais com maior nível de vídeo-realismo.

Em primeiro lugar, ressalta-se que o corpus audiovisual permitiu a captura de uma grande quantidade de imagens que, no contexto deste trabalho, não foram processadas. Considera-se importante a evolução dos algoritmos de processamento digital de imagens voltados para a detecção automática de elementos da face. Tal evolução permitirá que as imagens capturadas sejam processadas sem intervenção manual, possibilitando a análise mais aprofundada da dinâmica articulatória registrada por estas imagens.

Como decorrência da iniciativa de se obter algoritmos de análise das imagens mais eficientes, já foi desenvolvido um algoritmo de registro de imagens da base mais veloz e computacionalmente mais simples, baseado em transformações geométricas afins de translação, rotação e escalamento, como alternativa ao algoritmo baseado em funções de base radial apresentado na Seção 3.5.2. A aplicação de tal algoritmo também impacta a qualidade visual final da base de imagens e, desta forma, novos testes de inteligibilidade da fala estão planejados para avaliação desta influência.

Uma próxima etapa de desenvolvimento deste sistema está relacionada à busca de níveis de vídeo-realismo maiores, voltada para plataformas de maior desempenho, com recursos abundantes de memória e alta capacidade de processamento, através da implementação de uma base de imagens estendida. Nesta proposta, o aumento do vídeo-realismo será buscado através de um maior número de imagens armazenadas na base de imagens. Neste caso, a estratégia de síntese será baseada não somente na metamorfose entre visemas dependentes de contexto, mas também na concatenação e aproveitamento de imagens e pequenos fragmentos de vídeo originalmente gravados no corpus audiovisual. Partindo de uma base de imagens de tamanho mínimo (34 visemas), a base pode ser continuamente estendida procurando atender diferentes níveis de vídeo-realismo em função da capacidade da plataforma destino.

Outra característica importante a ser adicionada ao sistema é a capacidade de gerar animações com elementos de comunicação não verbais. Tais elementos podem ser adicionados à arquitetura já existente através da introdução de movimentações da cabeça e piscar de olhos.

Finalmente, a metodologia apresentada deve evoluir para ser capaz de sintetizar animações faciais adaptáveis a diferentes velocidades de discursos.

Referências Bibliográficas

AMIDROR, I. Scattered data interpolation methods for electronic imaging systems: a survey. *Journal of Electronic Imaging*, p. 157–176, 2002.

ARAD, N. et al. Image warping by radial basis functions: applications to facial expressions. *CVGIP: Graph. Models Image Process.*, Academic Press, Inc., Orlando, FL, USA, v. 56, n. 2, p. 161–172, 1994. ISSN 1049-9652.

BASU, S. et al. *Speech driven lip synthesis using viseme based hidden markov models*. Patente americana 6.366.885, requerida em 27 de Agosto de 1999 e concedida em 2 de Abril de 2002.

BENOÎT, C.; GOFF, B. L. Audio-visual speech synthesis from french text: Eight year of models, designs and evaluation at the ICP. *Speech Communication*, v. 26, n. (1-2), p. 117–129, October 1998.

BRADSKI, G.; KAEHLER, A. *Learning OpenCV: Computer Vision with the OpenCV Library*. [S.l.]: O'Reilly Media, 2008.

BRAND, M. Voice puppetry. In: *SIGGRAPH '99: Proceedings of the 26th Annual Conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999. p. 21–28. ISBN 0-201-48560-5.

BREGLER, C.; COVELL, M.; SLANEY, M. Video rewrite: driving visual speech with audio. In: *SIGGRAPH*. [S.l.: s.n.], 1997. p. 353–360.

BROOKE, N.; SCOTT, S. Two- and three-dimensional audio-visual speech synthesis. In: *Auditory-Visual Speech Processing*. [S.l.: s.n.], 1998. p. 213–218.

COHEN, M.; MASSARO, D. Modeling coarticulation in synthetic visual speech. In: *Computer Animation*. [s.n.], 1993. Disponível em: <citeseer.ist.psu.edu/cohen93modeling.html>.

COSATTO, E.; GRAF, H. P. Sample-based synthesis of photo-realistic talking heads. In: *CA*. [S.l.: s.n.], 1998. p. 103–110.

COSATTO, E.; GRAF, H. P. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia*, v. 2, n. 3, p. 152–163, 2000.

COSATTO, E.; GRAF, H. P.; HUANG, F. J. *System and method for triphone-based unit selection for visual speech synthesis*. Patente americana 7.209.882, requerida em 10 de Maio de 2002 e concedida em 24 de Abril de 2007.

COSATTO, E.; GRAF, H. P.; OSTERMANN, J. From audio-only to audio and video text-to-speech. *Acta Acustica united with Acustica*, v. 90, n. 6, p. 1084–1095(12), November/December 2004.

COSATTO, E. et al. Lifelike talking faces for interactive services. *Proceedings of the IEEE*, v. 91, n. 9, p. 1406–1429, September 2003.

COSTA, P. D. P.; DE MARTINO, J. M.; NAGLE, E. J. Sistema de animação facial 2D sincronizado com a fala integrado ao CPqD Texto Fala. In: *Anais do I Congresso Tecnológico Infobrasil - Infobrasil 2008*. Fortaleza, CE, Brasil: [s.n.], 2008.

CPQD. *Biblioteca CPqD Texto Fala - Manual de Utilização*. 2.7. ed. [S.l.], Maio 2007.

DE MARTINO, J. M. *Animação Facial Sincronizada com a Fala: Visemas Dependentes do Contexto Fonético*. Tese (Doutorado) — Universidade Estadual de Campinas, Julho 2005.

DE MARTINO, J. M.; VIOLARO, F. Benchmarking speech synchronized facial animation based on context-dependent visemes. In: *Proceedings of the 15th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2007*. [S.l.: s.n.], 2007. p. 105–112.

EDGE, J. D.; MADDOCK, S. Image-based talking heads using radial basis functions. In: *TPCG '03: Proceedings of the Theory and Practice of Computer Graphics 2003*. Washington, DC, USA: IEEE Computer Society, 2003. p. 74. ISBN 0-7695-1942-3.

EDGE M. SANCHES, S. M. J. Reusing motion data to animate visual speech. In: UNIVERSITY OF LEEDS. [S.l.]: AISB 2004 Convention Motion, Emotion and Cognition, 2004. p. 66–74.

EZZAT, T.; GEIGER, G.; POGGIO, T. Trainable videorealistic speech animation. In: *SIGGRAPH*. [S.l.: s.n.], 2002. p. 388–398.

EZZAT, T.; POGGIO, T. Miketalk: A talking facial display based on morphing visemes. In: *CA*. [s.n.], 1998. p. 96–102. Disponível em: <citeseer.ist.psu.edu/ezzat98miketalk.html>.

FARUQUIE, T. A. et al. Audio driven facial animation for audio-visual reality. In: *IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2001.

FISHER, C. G. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, n. 11, p. 796–804, 1968.

FOLEY, J. D. *Computer Graphics Principles and Practice*. Second. [S.l.]: Massachusetts: Addison-Wesley Publishing Company, 1990.

FORNEY, J. G. D. The Viterbi algorithm. *Proceedings of the IEEE*, v. 61, p. 268–278, March 1973.

GEIGER, G.; EZZAT, T.; POGGIO, T. *Perceptual Evaluation of Video-Realistic Speech*. Artificial Intelligence Laboratory, Massachussetss Institute of Technology, Cambridge, USA, February 2003. AI Memo 2003-03.

GOYAL, U. K.; KAPOOR, A.; KALRA, P. Text-to-audiovisual speech synthesizer. In: *Proceedings of the Second International Conference on Virtual Worlds*. [S.l.: s.n.], 2000. p. 256–269.

GRAF, H. P.; COSATTO, E.; EZZAT, T. Face analysis for the synthesis of photo-realistic talking heads. In: *IEEE International Conference on Automatic Face and Gesture Recognition*. [S.l.: s.n.], 2000. p. 189–194.

GRAF, H. P. et al. Visual prosody: facial movements accompanying speech. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2002. v. 2, p. 2037–2040.

GRATCH, J. et al. Creating interactive virtual humans: some assembly required. *Intelligent Systems, IEEE*, v. 17, n. 4, p. 54–63, Jul/Aug 2002. ISSN 1541-1672.

HORN, B. K.; SCHUNCK, B. G. *Determining Optical Flow*. Cambridge, MA, USA, 1980.

HUANG, F. J.; COSATTO, E.; GRAF, H. P. Triphone based unit selection for concatenative visual speech synthesis. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 2002. v. 2, p. 2037–2040.

INTERNATIONAL PHONETIC ASSOCIATION. *Handbook of the international phonetic association-a guide to the use of the international phonetic alphabet.* [S.l.]: Cambrige University Press, 1999.

ITU. *ITU-T Recommendation P.800: Methods for objective and subjective assessment of quality*. August 1996. Series P: Telephone Transmission Quality.

JÄHNE, B. *Digital Image Processing: Concepts, Algorithms, and Scientific Applications*. [S.l.]: Springer, 2005.

KSHIRSAGAR, S.; MAGNENAT-THALMANN, N. Visyllable based speech animation. *Comput. Graph. Forum*, v. 22, n. 3, p. 632–640, 2003. Disponível em: <http://dblp.uni-trier.de/db/journals/cgf/cgf22.html/KshirsagarM03>.

NG, K. *Survey of Data-Driven Approaches to Speech Synthesis*. 1998. Citeseer.ist.psu.edu/ng98survey.html.

NOH, J.; NEUMANN, U. *A Survey of Facial Modeling and Animation Techniques*. 1998.

PANDZIC, I. S.; OSTERMANN, J.; MILLEN, D. R. User evaluation: Synthetic talking faces for interactive services. *The Visual Computer*, Springer-Verlag, v. 15, n. 7-8, p. 330–340, November 1999. ISSN 0178-2789 (Print) 1432-2315 (Online).

PARKE, F. I. Computer generated animation of faces. In: *ACM'72: Proceedings of the ACM Annual Conference*. New York, NY, USA: ACM Press, 1972. p. 451–457.

PARKE, F. I.; WATERS, K. Computer facial animation. In: ____. Natick, MA, USA: A. K. Peters, Ltd., 1996. ISBN 1-56881-014-8.

PIGHIN, F. et al. Synthesizing realistic facial expressions from photographs. In: *SIGGRAPH Computer Graphics Proceedings*. [S.l.: s.n.], 1998. p. 75–84.

RUPRECHT, D.; MÜLLER, H. Image warping with scattered data interpolation. *IEEE Comput. Graph. Appl.*, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 15, n. 2, p. 37–43, 1995. ISSN 0272-1716.

SCOTT, K. C. et al. Synthesis of speaker facial movement to match selected speech sequences. In: *Proceedings of the Fifth Australian Conference on Speech Science and Technology*. [S.l.: s.n.], 1994. p. 620–625.

SHAFRANOVICH, Y. *RFC 4180 - Common Format and MIME Type for Comma-Separated Values (CSV) Files*. 2005. Internet Engineering Task Force.

SUMBY, W. H.; POLLACK, I. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, v. 26, n. 2, p. 212–215, March 1954.

TURING, A. M. Computing machinery and intelligence. *MIND*, v. 59, n. 236, p. 433–460, 1950.

UPPER Critical Values of the F Distribution. http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm. Página da Internet.

WOLBERG, G. Image morphing: a survey. *The Visual Computer*, v. 14, n. 8/9, p. 360–372, 1998.

WONNACOTT, T. H.; WONNACOTT, R. J. *Estatística Aplicada à Economia e à Administração*. [S.l.]: Livros Técnicos e Científicos Editora S.A., 1981.

ZITOVÁ, B.; FLUSSER, J. Image registration methods: a survey. *Image and Vision Computing*, v. 21, p. 977–1000, 2003.

Apêndice A

Logatomas e frases pronunciados para constituição do corpus audiovisual

A seguir são apresentados os itens pronunciados pela apresentadora durante processo de captura do corpus audiovisual descrito na Seção 3.2.1, do Capítulo 3.

O primeiro bloco é constituído de logatomas (palavras sem sentido) paroxítonos.

O segundo bloco é composto de três conjuntos de frases consideradas foneticamente ricas, onde a cada 9 frases ocorrem ao menos duas ocorrências de cada fonema do Português do Brasil.

•BLOCO 1 - LOGATOMAS

–Primeiro Grupo:

pipi, pipa, pipu, papi, papa, papu, pupi, pupa, pupu
fifi, fifa, fifu, fafi, fafa, fafu, fufi, fufa, fufu
titi, tita, titu, tati, tata, tatu, tuti, tuta, tutu
sissi, sissa, sissu, sassi, sassa, sassu, sussi, sussa, sussu
lili, lila, lilu, lali, lala, lalu, luli, lula, lulu
chichi, chicha, chichu, chachi, chacha, chachu, chuchi, chucha, chuchu
lhilhi, lhilha, lhilhu, lhalhi, lhalha, lhalhu, lhulhi, lhulha, lhulhu
kiki, kika, kiku, kaki, kaka, kaku, kuki, kuka, kuku
riri[1], rira, riru, rari, rara, raru, ruri, rura, ruru
rrirri[2], rrirra, rrirru, rrarri, rrarra, rrarru, rrurri, rrurra, rrurru

–Segundo Grupo:

ii, ia, iu
ei, ea, eu
éi, éa, éu
ai, aa, au

[1]Em “riri”, e variantes, a letra “r” é pronunciada fracamente como na palavra “carinho”.
[2]Em “rrirri”, e variantes, os dois “erres” são pronunciados fortemente como em “carrinho’.’

ói, óa, óu
oi, oa, ou
ui, ua, uu

–Terceiro Grupo:

dri, dra, dru
pri, pra, pru
pli, pla, plu
tri, tra, tru
tli, tla, tlu
cri, cra, cru
cli, cla, clu
fri, fra, fru
fli, fla, flu

•BLOCO 2 - FRASES FONETICAMENTE RICAS

–Conjunto 1

Não pode haver tréguas na guerra contra a malária.
Ninguém conhece um jeito melhor para escalar esta montanha.
Vários atletas tiveram um desempenho muito abaixo do esperado.
É a primeira vez que o nome do Brasil aparece nesta publicação.
As estatísticas oficiais mostram o avanço do desemprego na indústria.
O discurso de encerramento foi brilhante.
O inverno chegará bastante forte este ano.
A declaração do ministro sobre a flutuação do dólar é ambígua.
A carta de Pero Vaz de Caminha é o principal registro histórico do descobrimento.

–Conjunto 2

A partir de quinta-feira, os visitantes serão identificados através de um cartão magnético.
O conforto e a fartura das residências mais ricas contrasta com a pobreza da periferia.
Em qualquer lugar de Belém serve-se suco de açaí e água de coco.
O chefe do partido assumiu uma postura otimista durante a campanha eleitoral.
Nenhuma indústria pode viver permanentemente de subsídios.
A reforma econômica da China melhorou profundamente a gestão das empresas.
O suave perfume das gardênias é a principal característica do jardim central.
Um grande artista rejeita conselhos de curiosos.
Todos dizem que se trata do pior filme da história do festival de cinema.

–Conjunto 3

André evitou comentar suas falhas durante a atual crise bancária.

A ignorância impede o desenvolvimento tecnológico da região.
Trata-se de um episódio triste de nossa história.
Pedro Malan nega que irá defender a desvalorização cambial amanhã em Brasília.
É esta a grande interrogação que atravessa o livro.
As chuvas de junho atrapalharam a colheita do trigo.
Os filmes deste cineasta se distinguem pela direção precisa dos atores.
Não há tempo hábil para solucionar um problema tão complexo.
Todos acham que não existem crimes perfeitos onde a polícia é eficiente.

Apêndice B

Sistemas de Equações para Determinação dos Coeficientes das Funções de Base Radial

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1k} & 1 & x_1 & y_1 \\ \phi_{11} & \phi_{11} & \cdots & \phi_{11} & 1 & x_2 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \phi_{11} & \phi_{11} & \cdots & \phi_{11} & 1 & x_k & y_k \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 \\ x_1 & x_2 & \cdots & x_k & 0 & 0 & 0 \\ y_1 & y_2 & \cdots & y_k & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_k \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{B.1}$$

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1k} & 1 & x_1 & y_1 \\ \phi_{11} & \phi_{11} & \cdots & \phi_{11} & 1 & x_2 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \phi_{11} & \phi_{11} & \cdots & \phi_{11} & 1 & x_k & y_k \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 \\ x_1 & x_2 & \cdots & x_k & 0 & 0 & 0 \\ y_1 & y_2 & \cdots & y_k & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_k \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{B.2}$$

Onde:

•ϕ_{ij} - função radial calculada a partir dos pontos-âncora P_i e P_j:

$$\phi_{ij} = \sqrt{d(P_i, P_j)^2 + r_j^2} \tag{B.3}$$

•$d(P_i, P_j)$ - distância euclidiana entre os pontos P_i e P_j:

$$d(P_i, P_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{B.4}$$

•r_j é a distância do ponto-âncora P_j ao ponto-âncora vizinho mais próximo:

$$\begin{cases} r_j = min_{j \neq m}(d(P_j, P_m)) \\ j = 1...k \\ m = 1...k \end{cases} \tag{B.5}$$

•x_i e y_i são as coordenadas (x, y) dos pontos-âncora da imagem de entrada;

•x'_i e y'_i são as coordenadas (x, y) dos pontos-âncora da imagem de saída;

•α_i e β_i são os coeficientes da base de funções radiais (vide equações 3.1 e 3.2);

•a_i são os coeficientes do polinômio da função de interpolação (vide equações 3.2 e 3.4);

•$i, j = 1...k$.

Apêndice C

ANOVA Simples

A análise de variância, ou ANOVA (ANalysis Of Variance), visa avaliar se as diferenças observadas entre as médias de diferentes conjuntos de amostras podem ser consideradas estatiscamente significantes, ou seja, se elas são apenas consequência da variação amostral ou representam uma boa evidência da diferença entre as médias das populações.

Este problema é definido formalmente através do teste de hipóteses :

•H_0 (hipótese nula):$\mu_1 = \mu_2 = \mu_3 ... = \mu_r$ para r populações com n amostras (vide Tabela C.1).

•H_1: $\mu_i \neq \mu_k$ para algum par, $i \neq k$.

Tab. C.1: Conjuntos de Amostras

População	**Distribuição Suposta**	**Valores amostrais observados**
1	$N(\mu_1, \sigma^2)$	X_{1j} $(j = 1...n)$
2	$N(\mu_2, \sigma^2)$	X_{2j} $(j = 1...n)$
.	.	.
.	.	.
.	.	.
r	$N(\mu_r, \sigma^2)$	X_{rj} $(j = 1...n)$

A formulação simples (*one-way factor*) da ANOVA tem como objetivo a determinação do valor p que representa a probabilidade da hipótese nula ser verdadeira. Se o valor p é próximo a zero, isto sugere a rejeição desta hipótese significando que a média de ao menos um conjunto de amostras é significativamente diferente das outras médias.

Esta formulação é em geral expressa através da tabela de ANOVA que divide a variância dos dados em duas partes (Tabela C.2):

•variação devido a diferenças *entre* populações;

•variação devido a diferença entre os dados de um conjunto de amostras e sua média (variância *intergrupo*).

O cálculo das variâncias SMQ_r e SMQ_u da Tabela C.2, permite a obtenção da razão F, cuja distribuição é utilizada para testar H_0 no nível de 5%:

$$F = \frac{SMQ_r}{SMQ_u}$$

O valor de prova p pode então ser obtido através da obtenção dos valores críticos da distribuição de F (distribuição-F) levando-se em consideração os graus de liberdade das amostras e populações (UPPER...,).

Tab. C.2: Tabela de ANOVA (WONNACOTT; WONNACOTT, 1981)

Fonte de Variação	**Soma de Quadrados (SQ)**	**Graus de Liberdade**	**Variância**
Entre Populações	$n \sum_{i=1}^{r} (\overline{X_i} - \overline{\overline{X}})^2 = SQ_r$	$(r-1)$	$SMQ_r = \frac{SQ_r}{(r-1)}$
Intergrupo	$\sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \overline{X_i})^2 = SQ\mu$	$r(n-1)$	$SMQ_u = \frac{SQ_u}{r(n-1)}$