UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Silvano Ressurreição de Jesus Filho

Síntese de Animação Facial Expressiva baseada em Modelos Ocultos de Markov

Campinas

2021

Silvano Ressurreição de Jesus Filho

Síntese de Animação Facial Expressiva baseada em Modelos Ocultos de Markov

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Orientador: Profa. Dra. Paula Dornhofer Paro Costa

Este trabalho corresponde à versão final da dissertação defendida pelo aluno Silvano Ressurreição de Jesus Filho, e orientada pelo Profa. Dra. Paula Dornhofer Paro Costa.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

J499s

Jesus Filho, Silvano Ressurreição de, 1991-
Síntese de animação facial expressiva baseada em modelos ocultos de markov / Silvano Ressurreição de Jesus Filho. – Campinas, SP : [s.n.], 2021.

Orientador: Paula Dornhofer Paro Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Computação gráfica. 2. Animação por computador. 3. Expressão facial - Simulação por computador. 4. Modelos markovianos ocultos. I. Costa, Paula Dornhofer Paro, 1978-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: HMM-based expressive facial animation synthesis
Palavras-chave em inglês:
Computer graphics
Computer-based animation
Facial expression - Computer-based simulation
Hidden Markov Models
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Paula Dornhofer Paro Costa [Orientador]
Fernando Oscar Runstein
Tiago Fernandes Tavares
Data de defesa: 28-05-2021
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1813-2477
- Currículo Lattes do autor: http://lattes.cnpq.br/4951329740144076

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato(a): Silvano Ressurreição de Jesus Filho RA: 082796
Data de defesa: 24 de Maio de 2021
Título da Tese: "Síntese de Animação Facial Expressiva baseada em Modelos Ocultos de Markov"

Profa. Dra. Paula Dornhofer Paro Costa (Presidente)
Prof. Dr. Tiago Fernandes Tavares
Dr. Fernando Oscar Runstein

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

A Karina, que esteve sempre ao meu lado, mesmo nos momentos mais difíceis.

Aos meus irmãos Igor e Yêda, que me mostraram a direção, e me deram todo o suporte ao longo do caminho.

Especialmente à minha mãe, mulher capaz de coisas incríveis, que me ensinou e presenteou tudo que eu tenho de mais valioso.

Agradecimentos

Deixo meus agradecimentos à Fundação CPqD, em especial aos pesquisadores Flávio Olmos Simões e Mário Uliani Neto, pelas preciosas contribuições ao trabalho.

Agradeço ao apoio dos colegas das empresas em que trabalhei durante a realização desta pesquisa, Icaro Tech e Nubank.

Aos amigos que me acompanharam ao longo desse período, Felipe, Thiago, Guilherme, Arthur e William, que me mostrou o valor da garra quando o cenário é adverso.

Agradeço especialmente à minha orientadora, Profa. Dra. Paula Dornhofer Paro Costa, pela imensa dedicação e paciência, sem as quais este trabalho não seria possível. Seu exemplo me inspira a continuar trilhando a carreira acadêmica, mesmo nos tempos atuais, em que a Ciência não tem recebido o prestígio que merece.

Resumo

Agentes conversacionais virtuais, ou talking heads, são representações em vídeo da face de um agente virtual, que simulam a fala de um interlocutor humano. São uma poderosa ferramenta de interface computacional, com potencial para tornar interações mais naturais e atraentes. Nesse contexto, o campo de Síntese de Animação Facial lida com a geração automática de vídeos de um agente virtual, a partir de um texto arbitrário ou de áudio produzido previamente. Em sistemas de síntese de animação baseados em texto, o desafio técnico pode ser separado em duas etapas: geração de trajetórias de articulação labial e demais movimentações faciais a partir de uma sequência de fonemas, e conversão dessas trajetórias em sequências de imagens finais. Este trabalho foca na primeira etapa. Um sistema, baseado em modelos ocultos de Markov, capaz de gerar trajetórias de movimentações a partir de textos arbitrários é descrito e avaliado. A implementação é realizada para o Português do Brasil e a modelagem leva em consideração a expressividade do agente, sendo capaz de gerar trajetórias para diferentes emoções. A abordagem de modelagem de variancia do sinal é utilizada pra reduzir o efeito de sobreamortecimento.

Palavras-chaves: Computação Gráfica; Animação Facial; Modelos Ocultos de Markov

Abstract

Virtual conversational agents, or talking heads, are a powerful computer interface tool that simulates human speech, potentially making computer interactions more natural and engaging. The Visual Speech Synthesis field deals with automatic generation of videos of virtual agents, from generic text or audio. For text-based visual speech synthesis, the technical challenge can be split in two parts: the automatic generation of trajectories of attributes that describe visual speech from sequenes of phonemes, and the rendering of those trajectories into actual videos. This research focus on the first problem. We describe a Hidden Markov Models based system for the synthesis of visual attributes for Brazilian Portuguese. We also explore the inclusion of contextual expressive information and the modelling of variance to improve the quality of the generated trajectories.

Keywords: Computer Graphics. Facial Animation. Hidden Markov Models.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CPqD	Centro de Pesquisa e Desenvolvimento em Telecomunicações
HMM	*Hidden Markov Model*, ou Modelo Oculto de Markov
DNN	*Deep Neural Network*, ou Rede Neural Profunda
LSTM	*Long Short-Term Memory*

Lista de símbolos

$\boldsymbol{A}$	Matriz de probabilidade de transição entre estados
B	Conjunto das distribuições de probabilidade de emissão b_i
$\boldsymbol{F}$	Sequência temporal de atributos sonoros
$\boldsymbol{I}_n$	Matriz identidade de dimensão n
$\boldsymbol{K}$	Matriz de covariância
S	Conjunto de estados possíveis de um Modelo Oculto de Markov
$P(\cdot)$	Probabilidade
Q	Sequência temporal de estados
$\boldsymbol{V}$	Sequência temporal de atributos visuais
$\boldsymbol{Y}$	Sequência temporal da variável observável $\boldsymbol{y}_t$ de um Modelo Oculto de Markov
μ_V	Valor médito do vetor V
π_i	Probabilidade inicial do estado i

Sumário

1 Introdução

A evolução do campo de interfaces homem-máquina aponta para um futuro com formas mais naturais de interagir com dispositivos computacionais. Dispositivos de tecnologia vestível, aplicações de automação residencial, veículos autônomos e assistentes virtuais são prenúncios de uma realidade em que as interações serão menos dependentes de ferramentas como teclado e mouse, e serão mais próximas da maneira que os humanos interagem entre si e com o mundo físico. A busca por interações mais naturais alavanca o desenvolvimento de áreas de computação como Reconhecimento de Fala, Processamento de Linguagem Natural, Internet das Coisas e Inteligência Artificial.

Uma poderosa ferramenta de interface computacional são os agentes conversacionais virtuais, ou *talking heads*. *Talking heads* são representações em vídeo da face de um agente virtual, que simulam a articulação e fala de um interlocutor humano. São úteis em aplicações como assistentes virtuais, apresentadores de notícias, ensino e tutoria de idiomas, ou apoio à educação e terapias psiquiátricas (PANDZIC, 2002; COSATTO et al., 2003; DEY; MADDOCK; NICOLSON, 2010; MENDI; BAYRAK, 2013), tornando as interações computacionais mais naturais e atraentes.

Em casos de uso mais específicos, como assistentes virtuais de canais de atendimento, o roteiro das interações pode estar pré-definido. Nesses casos, uma técnica tipicamente empregada consiste em produzir previamente todos os vídeos das interações e utilizá-los ou combiná-los conforme a necessidade. Por outro lado, em aplicações mais gerais, como o exemplo apresentadores de notícias, prever antecipadamente todas as falas possíveis do agente pode ser infactível. A solução está em conseguir gerar automaticamente vídeos de um agente virtual a partir de textos arbitrários, processando, concatenando, gerando quadros estáticos da face do agente que tenham correspondência com a sequência de fonemas desejada, a fim de construir animações. Esse é o escopo do campo de Síntese de Animação Facial.

Evidentemente, este é um problema complexo. É necessário lançar mão de técnicas e algoritmos de aprendizado de máquina para tentar compreender e modelar a relação entre palavras e os movimentos articulatórios característicos do idioma em questão. Agentes virtuais realísticos devem dedicar atenção especial também a aspectos não-verbais da oralidade, como piscar de olhos, movimentos da face, sorrisos e demais manifestações de expressividade.

Um primeiro desafio técnico é a modelagem precisa do idioma. A busca aqui é por uma metodologia de modelagem da articulação capaz de converter uma sequência arbitrária de palavras em uma trajetória de movimentos articulatórios faciais realista.

Paralelamente, um segundo desafio diz respeito ao aspecto visual das imagens sintetizadas. É necessário gerar imagens faciais realistas a partir das movimentações articulatórias, para criar interações capazes de engajar o usuário final. Este trabalho foca no primeiro desafio.

Diversas estratégias já foram utilizadas para abordar esse problema. Implementações baseadas em Modelos Ocultos de Markov ou Redes Neurais Profundas seguem sendo o estado da arte na área (FAN et al., 2015; SATO; NOSE; ITO, 2017; WANG; SOONG, 2014). Esta dissertação parte das abordagens citadas e inclui à modelagem a dimensão expressiva, com o objetivo de alterar a trajetória sintetizada de acordo com a emoção desejada. O sistema proposto é implementado para uma base de vídeos em Português do Brasil.

As principais contribuições deste trabalho são:

- Implementação de um sistema baseado em Modelos Ocultos de Markov capaz de modelar o comportamento de atributos visuais a partir de uma sequência de fonemas em Português do Brasil. Até onde esta pesquisa pôde chegar, este é o primeiro trabalho do tipo para o idioma;
- Implementação e avaliação de abordagem para modulação expressiva da síntese de animação facial. A abordagem consiste em incluir rótulos de emoção à modelagem fonética.
- Implementação e avaliação de abordagem de modelagem de variância para a síntese de animação facial. Esta abordagem é frequentemente utilizada para a síntese de áudio e contribui para diminuir o efeito de *oversmoothing* (ver Seção 3.4.1), muito comum em sistemas envolvendo Modelos Ocultos de Markov. A avaliação objetiva mostra que a estratégia é eficaz para reduzir este efeito e melhorar a qualidade da trajetória sintetizada.

Esta dissertação é organizada da seguinte forma:

- O Capítulo 2 é uma revisão da literatura de Síntese de Animação Facial. O foco está em abordagens de síntese baseadas em modelos estatísticos, como os Modelos Ocultos de Markov. Também é feita uma breve introdução ao problema da modelagem de expressividade.
- O Capítulo 3 é uma descrição do sistema implementado neste trabalho, junto com os conceitos teóricos de Modelos Ocultos de Markov relacionados à sintese de animação facial. O capítulo começa com uma introdução à base de vídeos utilizada neste trabalho, e detalha o preprocessamento desta base para extrair informações no formato apropriado para a modelagem estatística. O preprocessamento aqui inclui extrações de atributos sonoros, visuais e transcrição fonética dos vídeos.

- O Capítulo 4 traz uma avaliação dos principais resultados do experimento. A avaliação objetiva mostra que o sistema é capaz de gerar uma trajetória de atributos que tem correlação com o resultado esperado. Uma inspeção visual subjetiva sugere que os resultados finais preservam conteúdo expressivo condizente com a emoção desejada.
- O Capítulo 5 apresenta as conclusões do trabalho e discute próximos passos para a pesquisa

2 REVISÃO DA LITERATURA

2.1 Introdução

O videorrealismo, ou seja, a capacidade de sintetizar animações capazes de serem confundidas com o vídeo de uma face real, representa o principal desafio no campo da síntese de Animação Facial. Interações com entidades artificiais (objetos, brinquedos, sistemas computacionais) podem se tornar pouco naturais, desconfortáveis e até repulsivas, e portanto inviáveis, se a representação visual não for suficientemente videorrealista (MORI; MACDORMAN; KAGEKI, 2012).

Um dos aspectos do videorrealismo é o fotorrealismo, que diz respeito à qualidade fotográfica da imagem da face. Uma imagem sintetizada será considerada fotorrealista se facilmente confundida com uma fotografia. No contexto da síntese de animações faciais, isso signica retratar com fidelidade proporções antropométricas e características como textura da pele, cabelo e rugas.

Uma animação videorrealista, além de fotorrealista, reproduz de forma convincente a dinâmica de movimentação e expressões faciais tipicamente associadas aos elementos verbais e não verbais característicos da fala humana. Em especial, uma animação videorrealista deve apresentar sincronia precisa do movimento dos lábios, levando em conta as articulações labiais características de cada idioma. Este detalhe é especialmente crítico, uma vez que a qualidade da sincronia labial da animação tem impacto significativo na inteligibilidade da fala (PANDZIC; OSTERMANN; MILLEN, 1999).

Nesse contexto, visema (palavra que tem origem na ideia de fonema visual) é uma postura articulatória visivelmente distinguível e que pode ser associada à produção de um ou mais fonemas. Via de regra, cada visema pode ser associado à produção acústica de mais de um fonema, como no caso dos fones [p] e [b], cujas produções envolvem posturas articulatórias visualmente muito semelhantes. Entretanto, um fonema específico pode estar associado a visemas diferentes, dependendo do contexto fonético em que é produzido, como o exemplo das palavras “para” e “pura”, onde o “p” muda de postura articulatória visivelmente influenciado pela vizinhança com o “a” ou o “u”. Esse efeito recebe o nome de coarticulação.

Na literatura, é possível identificar duas estratégias principais de representação da face: a abordagem baseada em modelos e a abordagem baseada em imagens.

A animação 3D parte de modelos tridimensionais para construir uma representação computacional da face humana. O realismo de modelos 3D depende do nível de detalhe

(a) Modelo facial 3D. Extraído de Bevacqua et al. (2007).

(b) Modelo facial 3D de alta resolução. Extraído de Cao et al. (2005)

(c) Modelo facial 2D. Extraído de Costa (2009)

Figura 2.1 – Exemplos de modelos faciais. Modelos faciais 2D são inerentemente fotorrealistas.

dedicado à simulação de textura e aspectos da pele. Os modelos mais avançados podem alcançar alto nível de qualidade, mas geralmente envolvem um alto custo computacional, e uma inspeção cuidadosa rapidamente revela a artificialidade da imagem.

Na abordagem de animação 2D, a animação facial é sintetizada através do processamento, sequenciamento, concatenação e apresentação de imagens de uma face real. Essa abordagem inerentemente busca o fotorrealismo, como observado na Figura 2.1c.

A maior vantagem da animação 3D é a liberdade de movimentação do agente conversacional, uma vez que, definida a representação do agente, quaisquer gestos e movimentos são facilmente impostos, enquanto que os mesmos movimentos, na animação 2D, dependem da existência de exemplos correspondentes na base de treinamento. A animação 2D, por outro lado, além de ser implementada com menor custo computacional, é mais fotorrealista que animação 3D. Por esse motivo, é mais adequada para simular os diálogos de uma pessoa real.

2.2 Síntese de Animação Facial 2D

O processo de síntese automática de animação facial consiste em gerar, ou selecionar de uma base de vídeos, uma sequência de imagens da face de um agente conversacional, como mostrado na Figura 2.2. Essa sequência de imagens deve estar em sincronia com a fala desejada e será utilizada para compor o video final da animação facial,

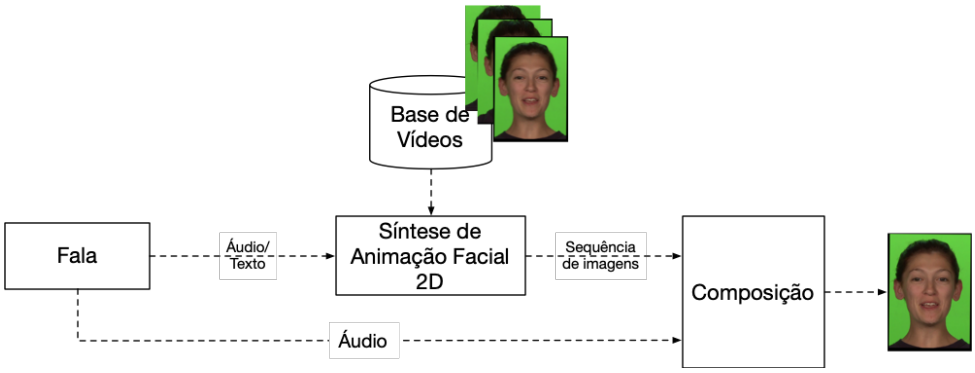


Figura 2.2 – Processo geral de síntese de animação facial 2D.

junto com o áudio correspondente.

Do ponto de vista computacional, entrada e saída do sistema da Figura 2.2 são representadas por sequências de símbolos ordenados temporalmente, aqui chamadas de variáveis sequenciais. A sequência de entrada pode ser áudio ou texto. No caso do texto, os símbolos em questão podem ser letras, fonenas, sílabas, etc. O áudio é comumente representado por coeficientes extraídos do espectro de potência do respectivo sinal auditivo.

O desafio técnico de sintetizar o vídeo de um agente conversacional virtual (em inglês, *talking head*) é essencialmente um problema de modelagem sequencial: gerar uma sequência de saída (imagens) a partir de uma sequência de entrada (fonemas ou áudio). Nesse aspecto, abordagens de treinamento e síntese de animação facial frequentemente guardam estreita similaridade com técnicas utilizadas no campo de síntese de áudio.

Historicamente, as várias abordagens de animação facial são separadas em três categorias principais:

- Metamorfose entre visemas
- Sistemas concatenativos
- Sistemas de síntese paramétrica

A estratégia de metamorfose entre visemas se baseia em selecionar quadros-chave, correspondentes aos fonemas a serem sintetizados. Para isso, é feita uma correspondência direta entre os fonemas da língua e quadros extraídos de uma base de vídeos. As transições entre os quadros-chave selecionados são geradas com alguma estratégia de interpolação de imagens. É similar à síntese de voz por seleção de unidades e é considerada a primeira técnica de síntese visual de fala (SCOTT et al., 1994).

Já os sistemas concatenativos sintetizam animações faciais concatenando trechos inteiros de vídeo de uma base de vídeos (BREGLER; COVELL; SLANEY, 1997; COSATTO; GRAF, 1998; Wesley Mattheyses; LATACZ; VERHELST, 2011). A premissa é a mesma

dos sistemas concatenativos de síntese de áudio (MOULINES et al., 1990): concatenar trechos reais de vídeo permite reproduzir a dinâmica dos movimentos articulatórios de forma mais realista que as transições da metamorfose entre visemas, que são criadas de forma artificial. O lado negativo desta estratégia é a necessidade de uma grande base de vídeos, que precisa satisfazer todas as combinações de fonemas possíveis, e depende da duração dos fragmentos armazenados.

Da mesma forma, sistemas baseados em síntese paramétrica também foram aplicados de maneira similar nos dois campos. Nesta classe de sistemas, os quadros da base de treinamento são representados por vetores de atributos visuais como, por exemplo, as posições de pontos-chave da face. A relação entre estes atributos visuais e a correspondente sequência de fonemas é modelada por alguma metodologia de predição estatística, com destaque para Modelos Ocultos de Markov e Redes Neurais Recorrentes.

A modelagem dos movimentos articulatórios através usando modelos de predição estatística alcança resultados convincentes, assim como os sistemas concatenativos, mas sem a necessidade de manter um grande volume de vídeos após o treinamento. Essa vantagem é crítica no caso de aplicações para dispositivos móveis, com menor capacidade de armazenamento e processamento. Além disso, a ascensão de técnicas de Aprendizado de Máquina e Inteligência Artificial trouxe os sistemas baseados em predição estatística para o primeiro plano no contexto das animações faciais. Os trabalhos recentes com melhores resultados se encaixam nesta categoria, e abordagens utilizando metamorfose entre visemas e sistemas concatenativos podem hoje ser considerados obsoletos no contexto da síntese de animação facial. Desta forma, esta revisão ficará restrita, daqui em diante, aos sistemas de síntese paramétrica, que serão discutidos mais profundamente na Seção 2.3.

2.3 Sistemas baseados em síntese paramétrica

Tornar o problema de animação facial apropriado à modelagem estatística envolve converter as imagens da base de vídeos para alguma representação numérica de atributos visuais. Por exemplo, pensando nos vídeos da base como uma sequência de quadros estáticos, cada quadro pode ser numericamente representado por um vetor com as posições geométricas de pontos-âncora da face, como mostrado na Figura 2.3. Tambem é comum utilizar Análise de Componentes Principais, ou outro algoritmo de redução de dimensionalidade, para transformar imagens em vetores numéricos (WANG; SOONG, 2014). A escolha de uma boa representação numérica é crucial para o realismo do vídeo final. A representação numérica ideal é capaz de parametrizar as imagens da base de vídeos de forma compacta e com perda mínima.

Sistemas de síntese paramétrica envolvem etapas de treinamento e síntese. Na fase de treinamento, o sistema é exposto a exemplos da base de vídeos, e tenta aprender o

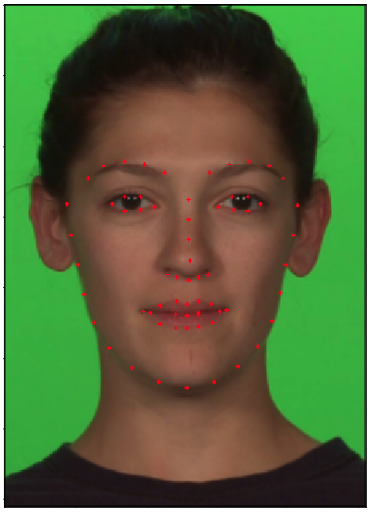


Figura 2.3 – Posição dos pontos âncoras da face (em vermelho). As coordenadas dos pontos-chave podem ser utilizadas como representação numérica das imgens

comportamento dos atributos visuais, a partir das variáveis de entrada. A base escolhida para treinamento deve ter uma boa variedade fonética e diversidade de contextos, para um treinamento mais robusto.

Na fase de síntese, o sistema tenta prever a trajetória dos atributos visuais de uma nova sequência de entrada. A trajetória sintetizada deve apresentar boa correspondência estática com a sequência de fonemas, mas tambem levar em conta o aspecto dinâmico, para evitar transições bruscas. A produção do vídeo final envolve ainda um passo de converter a trajetória de atributos de volta para imagens.

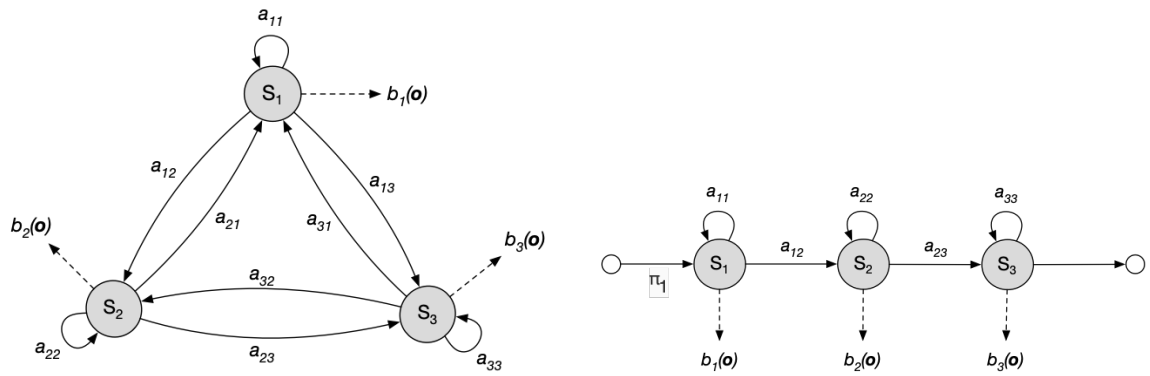
Até alguns anos atrás, trabalhos baseados em Modelos Ocultos de Markov eram a abordagem principal de síntese paramétrica (ver Subseção 2.3.1). Recentemente, a crescente disponibilidade de dados e poder de processamento alavancou o uso de redes neurais de arquitetura profunda (ver 2.3.2). Essa revisão foca nesses dois grupos de abordagens.

2.3.1 Modelos Ocultos de Markov

Nessa estratégia de modelagem baseada em Modelos Ocultos de Markov (do inglês *Hidden Markov Models*), a premissa é que o comportamento da variável sequencial aleatória $\boldsymbol{Y} = \{\boldsymbol{y}_t : t = 0, 1, \ldots\}$ no momento t, depende do estado q_t do sistema, que não pode ser diretamente observado. O objetivo é inferir características da sequência de estados $\boldsymbol{Q} = \{q_t : t = 0, 1, \ldots\}$, a partir de $\boldsymbol{Y}$.

Mais especificamente, um HMM é definido como uma máquina de estados finita, associada a um conjunto S de estados possíveis. Cada estado $S_i \in S$ está associado a uma distribuição de probabilidade de emissão $b_i[\boldsymbol{y}_t] = P(\boldsymbol{y} = \boldsymbol{y}_t | q_t = S_i)$ e as transições entre estados acontecem de acordo com a probabilidade $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, definida por

uma matriz de probabilidades de transição $\boldsymbol{A} = \{a_{i,j}\}$ (ver Figura 2.4)



(a) Modelo de Markov ergódico. As transições são possíveis entre todos os estados

(b) Modelo *left-to-right*. É a estrutura mais utilizada para a síntese de animação facial baseada em HMMs

Figura 2.4 – Exemplos de estruturas de HMMs.

A aplicação de HMMs se tornou difundida na área de reconhecimento de voz desde as décadas de 70 e 80, com os trabalhos seminais e com o desenvolvimento do sistema HTK (do inglês, Hidden Markov Model Toolkit) (JELINEK; BAHL; MERCER, 1975; BAKER, 1975; YOUNG, 1994). Posteriormente, o uso de HMMs se estendeu também à síntese de voz (TOKUDA; ZEN; BLACK, 2002; TODA; TOKUDA, 2005).

Em geral, nessas abordagens, a fala é representada por uma sequência de atributos extraídos do áudio. Unidades sonoras, como fonemas, sílabas ou palavras inteiras são representados por um ou mais estados ocultos. No campo de reconhecimento de voz, o problema é encontrar a sequência de estados que melhor se ajusta ao sinal observado, de acordo com a Equação 2.1. Já na área de síntese de voz, o desafio é o inverso: encontrar uma sequência de atributos de áudio que maximiza a Equação 2.2.

$$\boldsymbol{Q}^* = \arg\max P(\boldsymbol{Q}|\boldsymbol{Y}) \tag{2.1}$$

$$\boldsymbol{Y}^* = \arg\max P(\boldsymbol{Y}|\boldsymbol{Q}) \tag{2.2}$$

A formulação do problema de síntese de voz descrito acima abriu caminho para aplicação de HMMs também para a Síntese Visual de Fala. Essencialmente, o problema permanece idêntico, exceto pela inclusão de atributos visuais. Cada um dos estados está associado a uma distribuição de probabilidade gaussiana de emissão, e o processo de treinamento consiste, em linhas gerais, em encontrar os parâmetros dessa distribuição, encontrando a média e desvio padrão $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ que melhor se ajustam ao sinal observado. Embora não seja estritamente necessário, várias abordagens de síntese visual incluem também atributos acústicos na modelagem, para melhorar a sincronia labial do sinal sintetizado (TAMURA et al., 1999; WANG et al., 2010).

Na literatura de síntese visual, em geral são usados fonemas como unidades sonoras, e cada um dos fonemas é representado por três estados do HMM (TAMURA et al., 1998; YAMAMOTO; NAKAMURA; SHIKANO, 1998). É comum incluir nos fonemas informações sobre o contexto fonético, como os fonemas imediatamente anteriores e posteriores, tonicidade e posição relativa na sentença. Isso permite ao sistema HMM lidar com a questão de coarticulação.

Na fase de treinamento, é necessário que os dados estejam rotulados, com a transcrição dos fonemas correspondentes. Este é um ponto que dificulta a implementação de sistemas HMM para animação facial, uma vez que exige uma etapa custosa de segmentação fonética da base de treinamento, geralmente manual, e sujeita a erros.

Por outro lado, a estrutura de HMMs modela o sinal visual (e auditivo) diretamente. Em outras palavras, para determinado estado de um HMM, o sinal obtido na saída é um ajuste direto dos exemplos disponíveis na base de treinamento para o fonema correspondente, sem a necessidade de aprender relações matemáticas subliminares, como é o caso em redes neurais profundas, que discutiremos na próxima seção. Isto permite que HMMs aproximem o sinal visual observado, mesmo com poucos exemplos de treinamento. Wang e Soong (2014) apontam que 20 minutos de vídeo são suficientes pra treinar um sistema de animação facial baseado em HMMs.

2.3.2 DNNs

Em trabalhos recentes, modelos generativos baseados em redes com arquiteturas profundas, ou DNNs (do inglês, *Deep Neural Networks*) têm superado a performance de HMMs e se colocado como estado da arte na área de animação facial (FAN et al., 2015; TAYLOR et al., 2017; SHIMBA et al., 2015; SUWAJANAKORN; SEITZ; KEMELMACHER-SHLIZERMAN, 2017). Em linhas gerais, a estrutura técnica do problema a ser resolvido é similar ao caso da síntese HMM. Aqui, são utilizadas redes neurais recorrentes para modelar a trajetória de atributos visuais.

Redes neurais recorrentes são uma classe de redes neurais próprias para a modelagem de variáveis sequenciais. Redes desse tipo utilizam estados internos para armazenar informações recebidas anteriormente pela sequência de entrada. Isto permite processar sequências de entrada de tamanho variável e aprender aspectos da dinâmica temporal dos sinais envolvidos.

Especificamente, redes LSTM (do inglês, Long Short-Term Memory) são um tipo de rede recorrente que utiliza blocos de memória como estados internos, capazes de guardar informações durante um intervalo arbitrariamente longo (HOCHREITER; SCHMIDHUBER, 1997). Essa escolha de arquitetura permite às redes LSTM lidar com o problema do contexto na síntese de animação de forma mais geral. Enquanto a influência

do contexto fonético, na síntese HMM, é limitada pelas escolhas de formato e tamanho dos fonemas dependentes de contexto utilizados, as redes LSTM conseguem aprender quais informações tem mais influência no contexto futuro, e por quantas iterações essas informações devem ser mantidas.

Apesar das redes LSTM terem características promissoras para a modelagem da voz, inicialmente, os resultados conseguidos foram pouco impactantes, como apontam Graves, Mohamed e Hinton (2013). Os autores propuseram adicionar mais camadas ocultas de neurônios à arquitetura tradicional das redes LSTM. A implementação de redes LSTM com arquitetura profunda alcançou performance recorde em avaliações dentro do campo de reconhecimento de voz. Assim como aconteceu com HMMs, esse sucesso abriu caminho para aplicações em síntese de voz e síntese de animação facial (FAN et al., 2014; FAN et al., 2015).

Dentre as implementações direcionadas à síntese visual, em geral, são utilizadas versões bidirecionais de redes LSTM. Isto significa que as redes são capazes de aprender não somente com o contexto anterior, mas também com o contexto posterior, como mostrado na figura. Este detalhe é essencial para a modelagem visual da coarticulação, uma vez que a postura articulatória de determinado fonema é influenciada pelos fonemas seguintes, como no exemplo da letra “p” nas palavras “para” e “pura”.

Quanto à sequência de entrada, há uma tendência de modelar atributos de vídeo diretamente a partir do áudio. A intenção é dispensar uma etapa complexa de modelagem linguística, que inclui passos de transcrição e segmentação fonética da base de treinamento. Embora essa escolha torne o processo de treinamento mais simples, ela traz limitações que precisam ser tratadas. Shimba et al. (2015) enfrentaram problemas para modelar aspectos do contexto visual que não têm contrapartida no sinal auditivo, como o fato de que falantes moverem os lábios antes de produzirem som. Suwajanakorn, Seitz e Kemelmacher-Shlizerman (2017) apontam que, sem modelar explicitamente a expressão da fala, frequentemente a animação sintetizada pode parecer mais feliz ou mais triste que o esperado.

Fan et al. (2015) também destacam que, em comparação com a síntese HMM, o processo de treinar uma rede recorrente LSTM apresenta um custo computacional significativamente maior. Além disso, os trabalhos baseados em redes LSTM, em geral usam de uma base de vídeos de treinamento consideravelmente maior que trabalhos baseados HMM. Para efeito de comparação, Suwajanakorn, Seitz e Kemelmacher-Shlizerman (2017) utilizaram cerca de 17 horas de vídeo para treinar um rede LSTM, enquanto que Wang e Soong (2014) apontam que 20 minutos de vídeo são suficientes pra treinar um sistema de animação facial baseado em HMMs.

2.4 Expressividade

A percepção da naturalidade de um agente conversacional não se resume à sincronia da articulação labial com o conteúdo da fala. A expressividade é também uma dimensão importante da comunicação oral, e no contexto de síntese de animação facial, contribui para gerar interações mais convincentes (CHARFUELAN; STEINER, 2013).

A expressividade compreende aspectos visuais não-verbais relacionados à fala, como movimentações sutis da cabeça, olhar e posição das sobrancelhas. Estes sinais têm o poder de influenciar a receptividade do conteúdo transmitido, servir para enfatizar certa parte do discurso, e adicionar emoção à mensagem.

Naturalmente, animações faciais produzidas por sistemas de síntese paramétrica refletirão características de expressão presentes na base em que o sistema foi treinado. De fato, algumas implementações de animação facial impressionam pela capacidade de reproduzir trejeitos do sujeito presente nos vídeos originais, mesmo não tendo sido explicitamente construídas para tal (SUWAJANAKORN; SEITZ; KEMELMACHER-SHLIZERMAN, 2017). Entretanto, nesses casos, a expressão final do agente virtual pode não ficar compatível com a mensagem. Por exemplo, Suwajanakorn, Seitz e Kemelmacher-Shlizerman (2017) apontam que o vídeo sintetizado pode parecer muito sério para uma mensagem casual, e feliz demais para discursos mais sérios.

A modelagem explícita da emoção requer exemplos rotulados de vídeos com emoções específicas na base de treinamento. Nesse caso, o volume necessário de dados de treinamento cresce drasticamente com o domínio desejado de emoções. Em contrapartida, essa modelagem de emoção proporciona um grau adicional de controle sobre a expressividade da animação sintetizada.

A abordagem mais simples de incluir a dimensão da emoção na modelagem é treinar sistemas distintos para cada emoção. Nessa linha, Costa (2015) aplicou uma estratégia de metamorfose entre visemas para cada uma das 22 emoções descritas pelo modelo OCC, chegando a 782 visemas dependentes de contexto para o Português do Brasil (ORTONY; CLORE; COLLINS, 1990). Li et al. (2016) mostram que incluir vídeos com expressividade neutra no treinamento de cada um desses sistemas distintos pode contribuir com a performance e reduzir o volume necessário da base de treinamento.

Alternativamente, é possível usar todos os exemplos disponíveis na base para treinar um sistema unificado, introduzindo algum tipo de codificação emocional. Na síntese baseada em HMM's, isso significa trazer informações relacionadas à emoção para o contexto fonético (YAMAGISHI, 2005). Para a síntese baseada em redes recorrentes, basta incluir às entradas da rede atributos que codifiquem emoção (AN; LING; DAI, 2017). Em ambos os casos, é possível modelar não somente a categoria da emoção, como também sua intensidade, se houver disponível uma base de treinamento apropriadamente rotulada

(XUE; HAMADA; AKAGI, 2018).

2.5 Considerações Finais

Este capítulo apresentou um panorama histórico e do estado-da-arte da síntese de animação facial, com destaque para a síntese de animação facial expressiva.

Ainda que sistemas de síntese de animação facial baseados em algoritmos de modelagem *deep learning* estejam obtendo sucesso na síntese de sequências de imagens faciais expressivas fotorrealistas, existe o problema de como, idealmente, controlar a expressividade da animação sintetizada por meio de rótulos dados como entrada para o sistema.

Para atacar esse problema existe, primeiramente, o desafio de se obter bases audiovisuais de *talking heads* com rótulos de emoção ou expressividade, com o volume de dados necessário ao sucesso de treinamento de um modelo baseado em redes neurais profundas. Tal desafio é ainda mais pronunciado para línguas de menor interesse comercial (comparadas a línguas como o inglês e o mandarim).

Em segundo lugar, existe a discussão de se é possível estabelecer um padrão de expressividade na fala audiovisual, característica de um determinado rótulo emocional.

Para abordar o primeiro problema, o presente projeto adota a modelagem por HMMs (que requer menor volume de dados) e aproveita a disponibilidade da base audiovisual de fala expressiva CH-Unicamp. Dentre outras características (a serem detalhadas no capítulo seguinte) a base CH-Unicamp foi construída em condições controladas e traz amostras em português, rotuladas segundo o modelo de emoções de Ortony, Clore e Collins (modelo OCC) (ORTONY; CLORE; COLLINS, 1990).

O trabalho apresenta como contribuição principal a disponibilização de um modelo de síntese de animação facial expressiva baseado em HMMs para o português do Brasil, com controle de expressividade por rótulos emocionais.

Tal contribuição abre caminho para o estudo do segundo questionamento apontado: é possível encontrar (e reproduzir) um padrão na representação de expressividade de emoções e/ou diferentes estilos na fala audiovisual?

A existência de modelos de síntese de animação facial controlados por expressividade permite a elaboração de experimentos controlados de percepção subjetiva audiovisual, ao mesmo tempo contribuem para a avaliação dos modelos construídos. A realização de tais experimentos não foi o escopo deste trabalho, mas considera-se que o mesmo contribui de maneira importante para a longa jornada de compreensão dos mecanismos de percepção audiovisuais expressivos, com foco na língua portuguesa e sua cultura associada.

3 Método de Treinamento e Modelo de Síntese de Animação Facial Expressiva

Neste capítulo, apresentamos o método de treinamento e o modelo de síntese de animação facial expressiva propostos e implementados no contexto deste trabalho.

O método parte do pré-processamento de uma base de vídeos expressivos já existente para o português do Brasil. Após o processamento da base, os vídeos da base de treinamento são representados por atributos visuais e acústicos $\boldsymbol{V}_t$ e $\boldsymbol{F}_t$, respectivamente. Os atributos visuais são descritos por sequências de coeficientes PCA (do inglês, Principal Component Analysis) e os atributos acústicos descritos por coeficientes MFCC (do inglês, Mel-frequency Cepstral Coefficients). A extração de atributos dessa base é descrita na Seção 3.1.

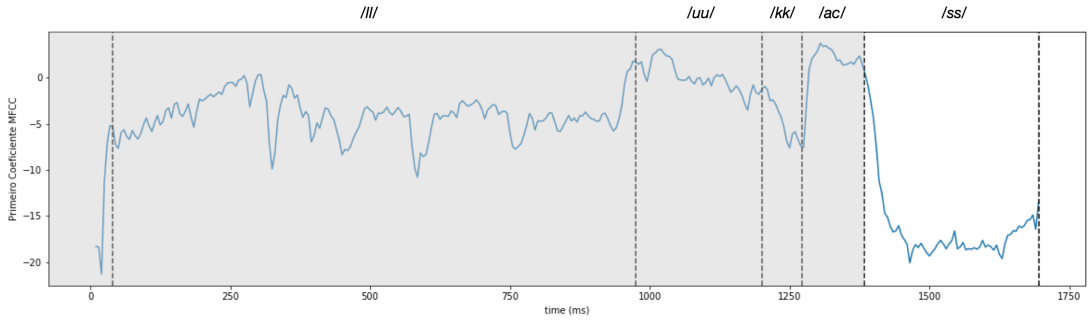
A Figura 3.1 ilustra como diferentes instâncias de um mesmo fone têm um perfil temporal característico, porém diverso, em diferentes produções acústicas, em diferentes contextos. A ideia geral por trás da modelagem da fala baseada em HMMs é treinar modelos de N estados específicos para cada fonema, a partir dos vários exemplos disponíveis desse fone na base de treinamento. Na fase de síntese, os HMMs treinados são concatenados, de acordo com a sequência de fonemas da frase desejada, buscando gerar uma trajetória de atributos, de acordo com o critério de Máxima Verossimilhança, a ser descrito na Seção 3.4.

Visando detalhar esse processo, a Seção 3.2 traz uma introdução dos conceitos e métodos relacionados ao treinamento e estimativa de parâmetros para Modelos Ocultos de Markov. Na Seção 3.3 são detalhados a estrutura de HMM usada neste trabalho para representar fonemas específicos e o passo-a-passo do processo geral de treinamento aplicado. No contexto do treinamento, atenção especial é dada à abordagem de modelagem do contexto fonético, exposta na Seção 3.3.3. Por fim, a Seção 3.4 mostra o processo de síntese de uma trajetória de atributos, a partir de uma sequência genérica de fonemas.

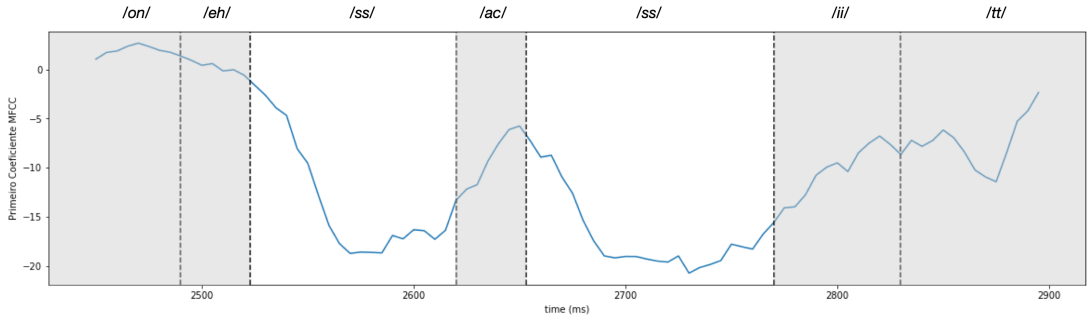
3.1 Extração de Atributos da Base de Dados de Treinamento

Como apresentado no capítulo anterior, Wang e Soong (2014) apontam que um sistema de animação facial baseado em síntese HMM necessita de uma base de treinamento de cerca de vinte minutos de vídeo de um falante.

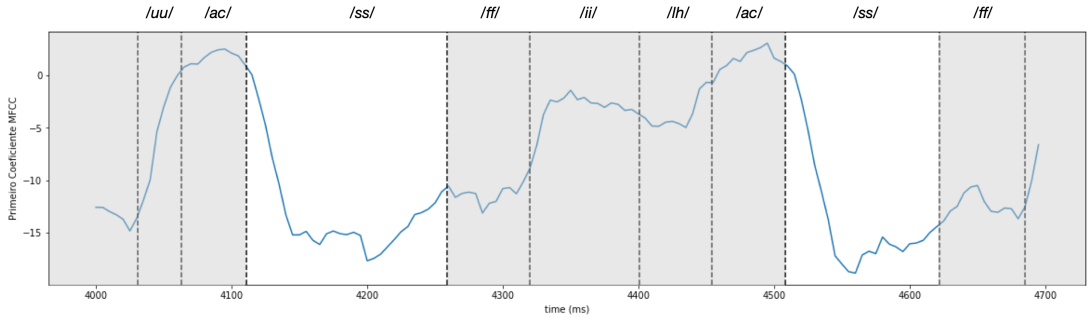
Os vídeos devem mostrar a face do falante em perspectiva frontal, com boa qualidade de imagem e iluminação. Deve-se tentar manter a movimentação e rotação da



(a) "*[...] Lucas [...]*"



(b) "*[...] com essa situação [...]*"



(c) "*[...] suas filhas [...]*"

Figura 3.1 – Comportamento dos atributos acústicos para diferentes instâncias do fonema */ss/*. Os exemplos tem perfil temporal similar.

cabeça ao mínimo possível. A base CH-Unicamp (ver Figura 3.2) compreende videos em alta resolução de uma atriz proferindo frases em língua portuguesa (COSTA, 2015). As frases foram selecionadas de forma a terem diversidade fonética, totalizando cerca de 25 minutos de vídeo.

Adicionalmente, cada uma das frases é produzida em duas versões: uma versão com expressividade neutra e uma outra versão de acordo com uma das 22 emoções descritas em Ortony, Clore e Collins (1990). O presente trabalho assume a hipótese de que um sistema de síntese pode aprender como cada uma dessas emoções influencia a expressividade da atriz, se fornecidos os dados correspondentes na etapa de treinamento.

O primeiro passo para a construção do sistema de síntese de animação facial é o pré-processamento da base, para converter os vídeos para o formato de dados esperado pelo

Figura 3.2 – Exemplo de quadro presente na base CH-Unicamp. As filmagens foram capturadas numa perspectiva frontal da face.

sistema. O pré-processamento de um vídeo específico consiste em transformar as dimensões visual e sonora para representações numéricas $\boldsymbol{V}$ e $\boldsymbol{F}$, respectivamente. Adicionalmente, estamos interessados também nos rótulos fonéticos temporizados dos vídeo de treinamento, uma vez que o processo de síntese HMM envolve treinar modelos distintos para cada fonema possível, baseado nos exemplos disponíveis na base de treinamento.

A Figura 3.3 sumariza o processo de extração de atributos dos vídeos da base. Ao longo das próximas seções, são detalhadas cada etapa do processo. Ao final desta etapa, obtém-se representações numéricas sincronizadas dos atributos visuais e atributos sonoros, além de transcrições fonéticas temporizadas dos vídeos da base de treinamento.

3.1.1 Extração de atributos de áudio

No campo de síntese HMM, é frequente utilizar coeficientes mel-cepstrais, ou MFCCs para representação de áudio (WANG; SOONG, 2014). MFCCs são coeficientes extraídos do espectro de potência de um trecho de sinal sonoro, tomando como referência a escala mel de frequência (TIWARI, 2009). A vantagem dos MFCCs em relação a outras estratégias de representação é aproximar melhor as faixas de frequência percebidas pela audição humana.

Na configuração utilizada neste trabalho, vetores $\boldsymbol{f}_t$ com 12 coeficientes mel-cepstrais são extraídos em janelas de $20ms$, com deslocamento de $5ms$ entre duas janelas consecutivas. Nesse caso, um trecho de áudio de $1s$, é convertido para uma sequência $\boldsymbol{f}$ de dimensões 12×197.

Além disso, também é comum incluir à modelagem, os atributos dinâmicos $\Delta\boldsymbol{f}_t$ e $\Delta\Delta\boldsymbol{f}_t$, respectivamente as derivadas de primeira e segunda ordem do sinal acústico. A informação dinâmica adiciona suavidade ao sinal sintetizado, tornando o resultado final mais fluido e natural.

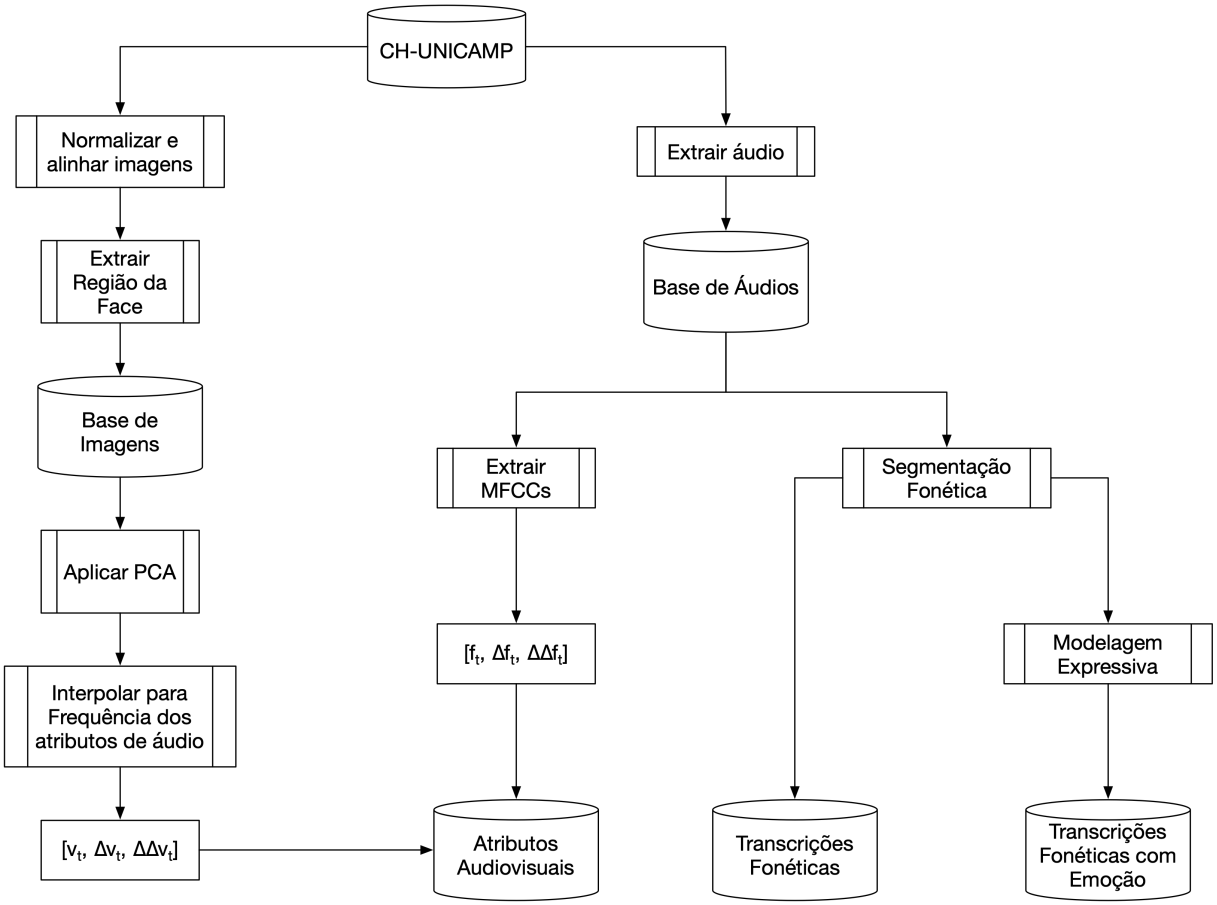


Figura 3.3 – Visão geral do pré-processamento.

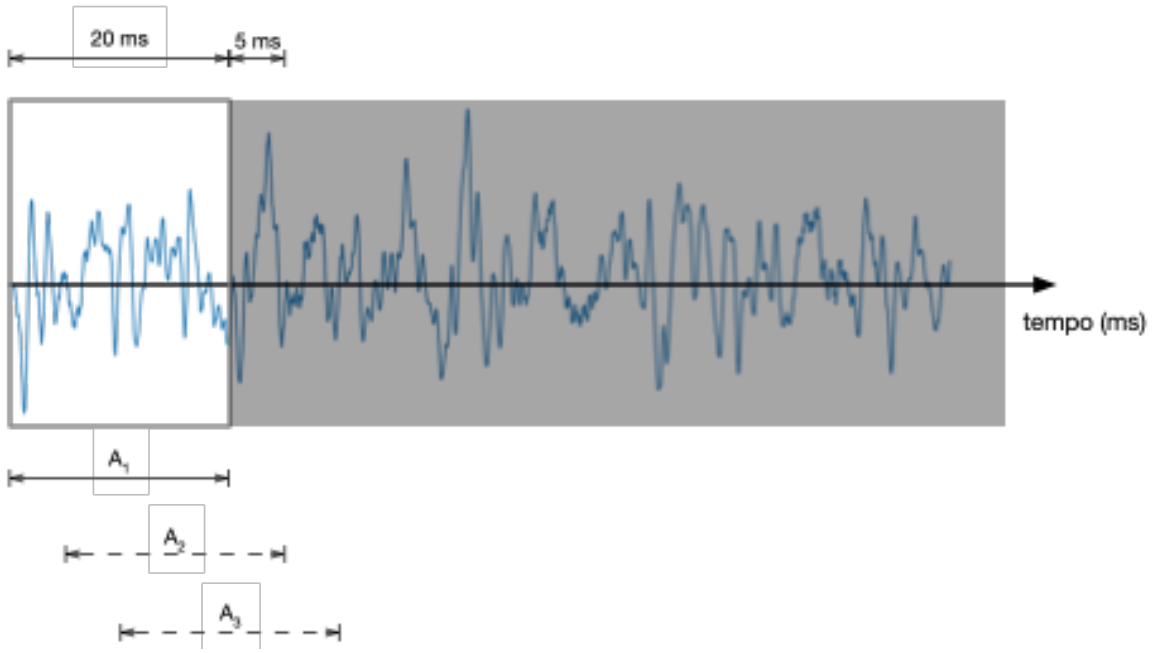
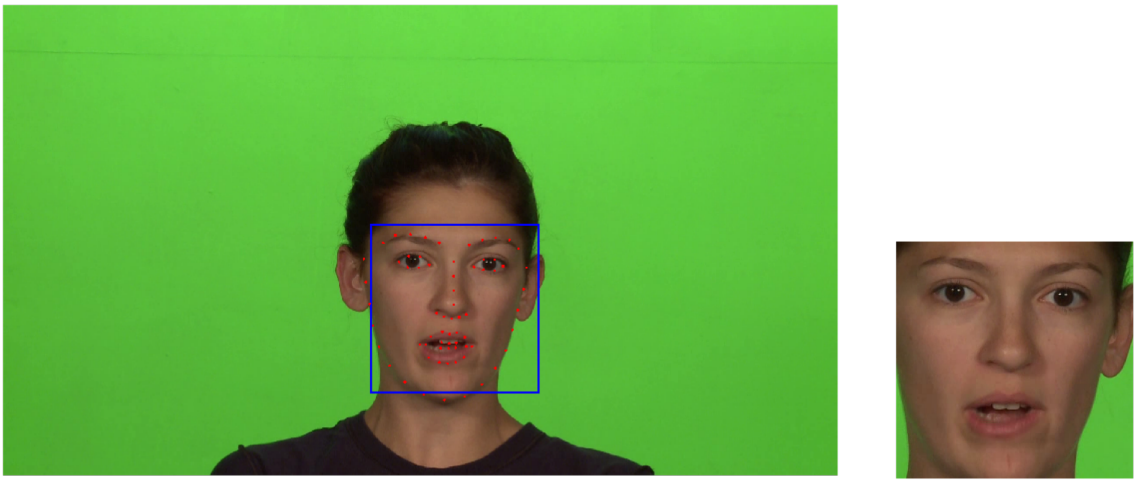


Figura 3.4 – Extração de atributos a partir da onda sonora. Os coeficientes MFCC são extraídos em janelas de 20 ms, com deslocamentos de 5ms.

3.1.2 Extração de atributos visuais

Visualmente, o objetivo do sistema é modelar os movimentos da face do agente no vídeo. Isso inclui movimentação labial e sinais não verbais, como posição da bochecha e dos olhos. A etapa de pré-processamento aqui consiste em extrair de cada quadro dos vídeos um destaque da face do agente. Como o agente do vídeo pode naturalmente mover ou rotacionar a cabeça, é necesssário normalizar as imagens para uma visão frontal e alinhada da face.

O modelo proposto por Kazemi e Sullivan (2014) é capaz de detectar 68 pontos de referência na face, chamados pontos-chave, com alta confiabilidade. A normalização das imagens é feita tomando como referência esses pontos-chave, como mostra Figura 3.5. Esta etapa inclui transformações de translação e rotacão para um alinhamento vertical da face, e alterar a escala das imagens para que todas tenham as mesmas dimensões finais. Cootes et al. (1995) propõem um método iterativo para transformar as imagens iniciais em direção a um alinhamento ótimo dos pontos-chave.



(a) Quadro original

(b) Destaque da face após o processo de normalização

Figura 3.5 – Normalização dos quadros. O pré-processamento da base de vídeos inclui destacar a região da face em todos os quadros

O método PCA permite converter as imagens alinhadas para uma representação numérica compacta. O vetor $\boldsymbol{v}_t$ é formado com os primeiros 20 componentes PCA, que explicam 91% da variância. Como os vídeos da base foram capturados a uma taxa de 30 quadros por segundo, é necessário fazer uma reamostragem temporal para que os atributos visuais estejam sincronizados com os atributos sonoros. Por fim, assim como nos atributos de áudio, são extraídos também os componentes dinâmicos $\Delta\boldsymbol{v}_t$ e $\Delta\Delta\boldsymbol{v}_t$.

3.1.3 Segmentação fonética

O requisito inicial para obter a transcrição fonética segmentada dos vídeos de treinamento é a representação fonética para o idioma. A representação fonética é a correspondência entre os fones e os símbolos linguísticos respectivos, chamados de fonemas. Neste trabalho, utilizamos a representação fonética desenvolvida e gentilmente disponibilizada pela Fundação CPqD - Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPQD), e representada no Apêndice A.

O processo de segmentação fonética consiste em identificar, nos videos de treinamento, as fronteiras temporais entre as ocorrências de diferentes fonemas, como mostra a Figura 3.6 A qualidade da segmentação pode influenciar o sucesso de posteriores aplicações. Por isso, é desejado que as fronteiras sejam corretamente identificadas com precisão de milissegundos.

A segmentação manual dos arquivos é um processo demorado, repetitivo, e sujeito a erro humano. Alternativamente, há formas de realizar a segmentação automaticamente. Neste trabalho, segmentação fonética foi realizada com o auxílio do pacote Kaldi para reconhecimento de fala, usando a arquitetura GMM/HMM (POVEY et al., 2011). O processo parte dos atributos acústicos definidos na Seção 3.1.1 para chegar numa transcrição fonética temporizada, indicando as fronteiras entre diferentes fonemas, como mostrado na Figura 3.6.

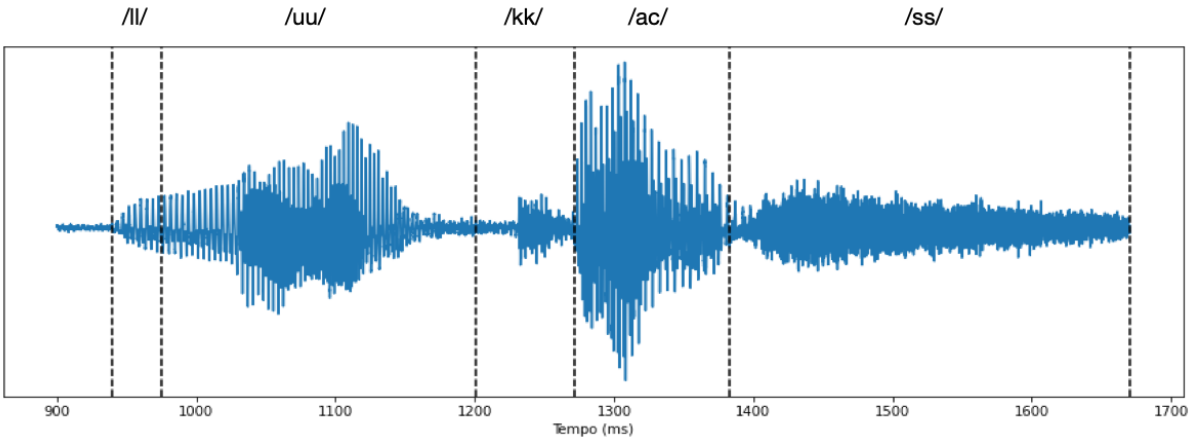


Figura 3.6 – Segmentação fonética. Diferentes fonemas tem padrões distintos de forma de onda.

Um ajuste fino foi aplicado às posições das fronteiras obtidas pela segmentação, a partir da rede neural convolucional proposta por Cuozzo et al. (2018).

A fim de modelar os efeitos de coarticulação, é necessário utilizar informações do contexto fonético. É comum utilizar incluir os fonemas vizinhos na definição do contexto, transformando os fonemas em trifones ou pentafones. Tonicidade e posição relativa do fonema na sentença também adicionam informações importantes na modelagem. A definição

Início [100ns]	Fonema	Fonema Dependente de Contexto
201250	#p	yy^yy-#p+ll=uu/P1:yy/P2:yy/P3:yy/P4:yy/P5:yy/P6:yy/P7:yy/P8:yyyy/P9:yy/P10:00
402500	ll	yy^#p-ll+uu=kk/P1:-1/P2:+0/P3:-4/P4:01/P5:00/P6:03/P7:00/P8:yyy0/P9:00/P10:00
9753750	uu	#p^ll-uu+kk=ac/P1:+0/P2:+1/P3:-3/P4:01/P5:00/P6:03/P7:00/P8:yyy0/P9:00/P10:00
12018125	kk	ll^uu-kk+ac=ss/P1:+1/P2:+2/P3:-2/P4:04/P5:00/P6:03/P7:00/P8:yyy0/P9:00/P10:00
12726250	ac	uu^kk-ac+ss=#e/P1:+2/P2:+3/P3:-1/P4:04/P5:00/P6:03/P7:00/P8:yyy0/P9:00/P10:00
13835624	ss	kk^ac-ss+#e=tt/P1:+3/P2:+4/P3:+0/P4:04/P5:00/P6:03/P7:00/P8:yyy0/P9:00/P10:00
16964374	#e	ac^ss-#e+tt=uu/P1:yy/P2:yy/P3:yy/P4:yy/P5:yy/P6:yy/P7:yy/P8:yyyy/P9:yy/P10:00

Tabela 3.1 – Exemplo final de transcrição fonética segmentada, para fonemas isolados, ou fonemas com o contexto fonético.

do contexto fonético utilizada neste trabalho, fornecida pela Fundação CPqD, é especificada no Apêndice B.

A Tabela 3.1 mostra um exemplo do resultado final do processo de segmentação, tanto para fonemas isolados, quanto para fonemas dependentes de contexto. Ambas as representações são importantes e são utilizadas no processo de treinamento.

3.1.4 Expressividade

A modelagem da expressividade requer que a informação das emoções correspondentes ao vídeos seja disponibilizada explicitamente na etapa de treinamento. De outra forma, a animação final simplesmente absorveria características visuais dos exemplos de treinamento, mas sem guardar necessariamente uma correspondência consistente com o conteúdo da fala do agente.

A base CH-Unicamp já disponibiliza uma correspondência entre os vídeos e as respectivas emoções baseada no modelo de Ortony, Clore e Collins (1990). Incluir ao contexto fonético um código que identifique a emoção correspondente a cada vídeo (ver Tabela B.1) garante que o mesmo fonema, proferido com duas emoções diferentes, seja tratado como dois fonemas distintos no processo de treinamento. Dessa forma, numa posterior etapa de síntese, especificar a emoção da animação a ser sintetizada garante que sejam utilizados apenas exemplos de fonemas que tenham relação com a emoção desejada.

3.2 Modelagem HMM: Conceitos Relacionados

Um modelo oculto de Markov é definido pelos seguintes elementos:

1. Um conjunto de N estados possíveis $S = \{S_i\}$, para $i = 1, \ldots, N$. Aqui, uma sequência particular de estados de comprimento T será denominada $\boldsymbol{Q} = \{q_t\}$, onde $t = 1, \ldots, T$ representa a dimensão temporal e $q_t \in S$.

2. Uma distribuição de probabilidade de estados inicial $\boldsymbol{\Pi} = \{\pi_i\}$, onde

$$\pi_i = P(q_1 = S_i), \quad i = 1, \ldots, N \tag{3.1}$$

3. Uma matriz de probabilidade de transição entre estados $\boldsymbol{A} = \{a_{ij}\}$, onde

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad i, j = 1, \ldots, N \tag{3.2}$$

4. Distribuições $B = \{b_i\}$ de probabilidade de emissão da variável observável $\boldsymbol{Y} = \{y_i\}$, onde

$$b_i(y_t) = P(y = y_t | q_t = S_i) = N(\boldsymbol{\mu}_i, \boldsymbol{K}_i), \quad i = 1, \ldots, N \tag{3.3}$$

Aqui serão utilizadas distribuições gaussianas como probabilidades de emissão. As variáveis $\boldsymbol{\mu}_i$ e $\boldsymbol{K}_i$ representam, respectivamente, o vetor de valores médios e a matriz de covariância da distribuição de probabilidade correspondente ao estado

O conjunto dos parâmetros que define um HMM é descrito por $\lambda = (\boldsymbol{A}, B, \boldsymbol{\Pi})$.

3.2.1 Sequência Ótima de Estados

Uma problema comum na modelagem HMM é encontrar uma sequência de estados ótima $\boldsymbol{Q}^* = \{q_t^*\}$, dada uma sequência de saída $\boldsymbol{Y} = \{y_t\}$. Uma aproximação frequente é encontrar o caminho de máxima verossimilhança, ou seja:

$$\boldsymbol{Q}^* = \underset{\boldsymbol{Q}}{\operatorname{argmax}}\, P(\boldsymbol{Y}, \boldsymbol{Q}) \tag{3.4}$$

Essa sequência ótima pode ser obtida pelo algoritmo de Viterbi, descrito a seguir, onde $\psi_t(i)$ é a sequência de estados mais provável que termine com estado i no tempo t e $\delta_t(i)$ é a probabilidade dessa sequência (YAMAGISHI, 2006).

1. Inicialização

$$\delta_1(i) = \pi_i b_i(y_1), \tag{3.5}$$
$$\psi_1(i) = 0, \tag{3.6}$$
$$\text{para } i = 1, \ldots, N$$

2. Recursão

$$\delta_t(j) = \max_i(\delta_{t-1}(i) a_{ij}) y_t, \tag{3.7}$$
$$\psi_t(j) = \underset{i}{\operatorname{argmax}}(\delta_{t-1}(i) a_{ij}), \tag{3.8}$$
$$\text{para: } \begin{matrix} i = 1, \ldots, N, \\ t = 2, \ldots, T \end{matrix}$$

3. Fim

$$P(\boldsymbol{Y}, \boldsymbol{Q}^*) = \max_i(\delta_T(i)), \tag{3.9}$$

$$q_T^* = \text{argmax}(\delta_T(i)), \tag{3.10}$$

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \tag{3.11}$$

$$\text{para } t = 2, \ldots, T$$

3.2.2 Estimativa de Parâmetros

O problema de encontrar um conjunto de parâmetros λ^* que maximiza a probabilidade $P(\boldsymbol{Y}|\lambda)$ não tem solução analítica conhecida. É necessário fazer uso de abordagens iterativas para atualizar os parâmetros λ, aproximando-os de λ^*. Duas estratégias principais geralmente são utilizadas no treinamento de HMMs: o método de Máxima Verossimilhança e o algoritmo de Expectativa-Maximização (YAMAGISHI, 2006).

Máxima Verossimilhança

Na abordagem de máxima verossimilhança, os parâmetros λ são atualizados de acordo com a sequência ótima de estados, encontrada com o auxílio do algoritmo de Viterbi (ver 3.2.1). Para distribuições de emissão gaussianas, o algoritmo é descrito a seguir:

1. A partir dos valores atuais dos parâmetros λ, encontrar a sequência de estados ótima, de acordo com o algoritmo de Viterbi. Nesse caso, para cada estado S_i, $C_i = \{t|q_t = i\}$ representa o conjunto de todos os instantes t ocupados pelo estado i, e c_i a quantidade de elementos de C_i

2. Para essa sequência de estados recalcular os parâmetros do HMM como:

$$\boldsymbol{\mu}_i = \frac{\sum\limits_{t \in C_i} \boldsymbol{y}_t}{c_i} \tag{3.12}$$

$$\boldsymbol{K}_i = \frac{\sum\limits_{t \in C_i} (\boldsymbol{y}_t - \mu_i)(\boldsymbol{y}_t - \mu_i)^T}{c_i} \tag{3.13}$$

$$a_{ij} = \frac{|\{t|q_t = i, q_{t+1} = j\}|}{|\{t|q_t = i\}|} \tag{3.14}$$

3. Repetir até que a convergência seja atingida.

A abordagem de máxima verossimilhança é também chamada de treinamento Viterbi.

Expectativa-Maximização

O algoritmo de expectativa-maximização, também conhecido como algoritmo Baum-Welch, é semelhante ao treinamento Viterbi. A principal diferença é que o algoritmo Baum-Welch é uma otimização mais local e sensível dos parâmetros do HMM. Aqui, em cada iteração, calculamos a probabilidade dos dados observados e, a partir desse valor, os parâmetros são otimizados:

1. A partir de um conjunto inicial de parâmetros λ, encontrar as probabilidades γ e ξ, sendo $\gamma_t(i)$ a probabilidade de o tempo t ser ocupado pelo estado i e $\xi_t(i,j)$ a probabilidade conjunta de o tempo t ser ocupado pelo estado i e o tempo seguinte $t+1$ ser ocupado pelo estado j:

$$\gamma_t(i) = P(\boldsymbol{Y}, q_t = i|\lambda), \tag{3.15}$$

$$\xi_t(i,j) = P(\boldsymbol{Y}, q_t = i, q_{t+1} = j|\lambda) \tag{3.16}$$

2. Recalcular os parâmetros λ do HMM, a partir das probabilidades acima, como:

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{3.17}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) y_t}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{3.18}$$

$$\boldsymbol{K}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(\boldsymbol{y}_t - \mu_i)(\boldsymbol{y}_t - \mu_i)^T}{\sum_{t=1}^{T} \gamma_t(i)} \tag{3.19}$$

3. Repetir até que a convergência seja atingida.

3.3 Treinamento

Neste trabalho, cada fonema é representado por um HMM *left-to-right* de 5 estados, como mostrado na Figura 3.7 Em estruturas *left-to-right*, os estados do HMM estão dispostos em uma sequência definida, e as transições só podem acontecer de um estado para si próprio ou para um único estado seguinte. Cada um dos estados produz observações dos vetores de atributos de vídeo e áudio, V_t e A_t respectivamente, de acordo com distribuições gaussianas multidimensionais.

$$p(\boldsymbol{V}_t|q_t = S_i) = N(\boldsymbol{\mu}_i^{(\boldsymbol{V})}, \boldsymbol{K}_i^{(\boldsymbol{V})}) \tag{3.20}$$

$$p(\boldsymbol{F}_t|q_t = S_i) = N(\boldsymbol{\mu}_i^{(\boldsymbol{F})}, \boldsymbol{K}_i^{(\boldsymbol{F})}) \tag{3.21}$$

A estrutura de HMM descrita, depois de treinada, é capaz de sintetizar não só o sinal de vídeo, mas tambem o sinal de áudio da fala. Na prática, entretanto, a síntese de áudio necessita de um volume consideravelmente maior de dados de treinamento para um resultado final satisfatório. A decisão de incluir atributos de áudio e vídeo conjuntamente se justifica pelo fato de os dois sinais serem altamente correlacionados, no contexto da fala. Isso significa que o sinal de áudio pode ajudar no processo de detecção de alinhamento de estados, bem como na qualidade da sincronia labial do video final sintetizado, mesmo que não estejamos especialmente interessados em sintetizar áudio, nesse caso.

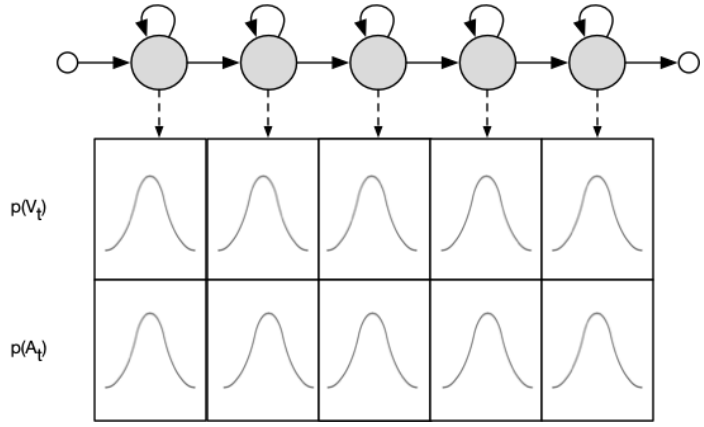


Figura 3.7 – HMM *left-to-right*. Cada um dos estados produz observações dos vetores de atributos de vídeo e áudio, V_t e A_t respectivamente, de acordo com distribuições gaussianas multidimensionais

Os vários HMM correspondentes aos fonemas são inicializados com os mesmos parêmtros. Os parâmetros de cada HMM são atualizados de acordo com o processo da Figura 3.8, a partir dos dados de treinamento. Ao fim do processo, temos modelos específicos para cada fonema, que podem ser aplicados num processo de síntese posterior. A seguir, cada uma das etapas é descrita com mais detalhes.

3.3.1 Inicialização

Os modelos correspondentes a cada um dos fonemas são inicializados com parâmetros genéricos, mostrados na Equação 3.22. A partir daí, cada um dos modelos é treinado, primeiramente com o treinamento Viterbi, em seguida de acordo com o algoritmo Baum-Welch. Em cada iteração, os modelos são atualizados com base em todos os exemplos disponíveis do fonema correspondente na base de treinamento.

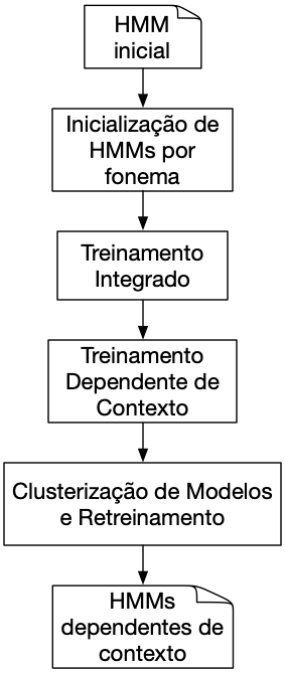


Figura 3.8 – Processo geral de treinamento de HMMs

$$\pi_i = \begin{cases} 1, & \text{se } i = 1 \\ 0, & \text{caso contrário} \end{cases}$$

$$a_{ij} = \begin{cases} 0.6, & \text{se } j = i \\ 0.4, & \text{se } j = i + 1 \\ 0, & \text{caso contrário} \end{cases} \tag{3.22}$$

$$\boldsymbol{\mu}_i = \vec{0}$$

$$\boldsymbol{K}_i = I_n$$

3.3.2 Treinamento Integrado

Até aqui, o processo de treinamento foi realizado sobre modelos isolados, correspondendo a fonemas específicos. Apesar de ser suficiente para a síntese de fala, a principal etapa da construção de modelos HMM gira em torno do conceito de treinamento integrado.

No treinamento integrado, os parâmetros de vários modelos diferentes são atualizados conjuntamente. Aqui, a ideia é concatenar os modelos isolados para compor HMMs que correspondam a exemplos da base de treinamento, como ilustra a Figura 3.9. Assim, o treinamento é realizado de acordo com o algoritmo Baum-Welch, a partir de trechos e frases inteiras. A principal vantagem é permitir uma sintonia mais fina dos parâmetros nas regiões de fronteira entre os fonemas, uma vez que o processo de

segmentação fonética está sujeito a falhas.

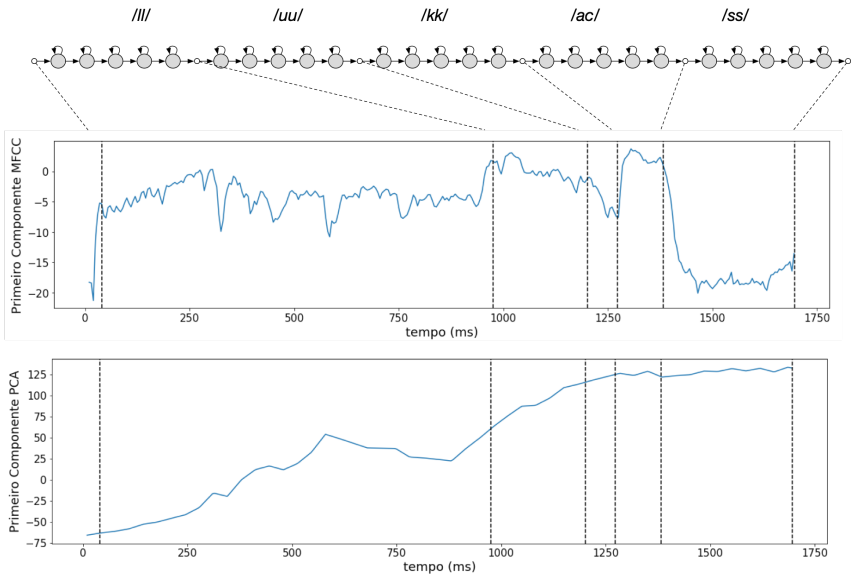


Figura 3.9 – Concatenação de HMMs isoladas para compor o modelo correspondente à palavra "Lucas"

3.3.3 Modelos Dependentes de Contexto

A estratégia para modelar a coarticulação na síntese HMM é utilizar HMMs dependentes de contexto. Na prática, isso significa utilizar modelos distintos para descrever o comportamento de um mesmo fonema, quando encontrado em diferentes contextos fonéticos.

Os modelos treinados até aqui, correspondentes a fonemas isolados, servem como ponto de partida para os modelos dependentes de contexto. Cada um dos modelos é copiado para as várias possibilidades de modelos dependentes de contexto, como ilustrado na Figura 3.10, e o treinamento integrado é repetido. Agora, as diferenças entre modelos diferentes de um mesmo fonema refletem a influência do contexto fonético nos atributos acústicos e visuais.

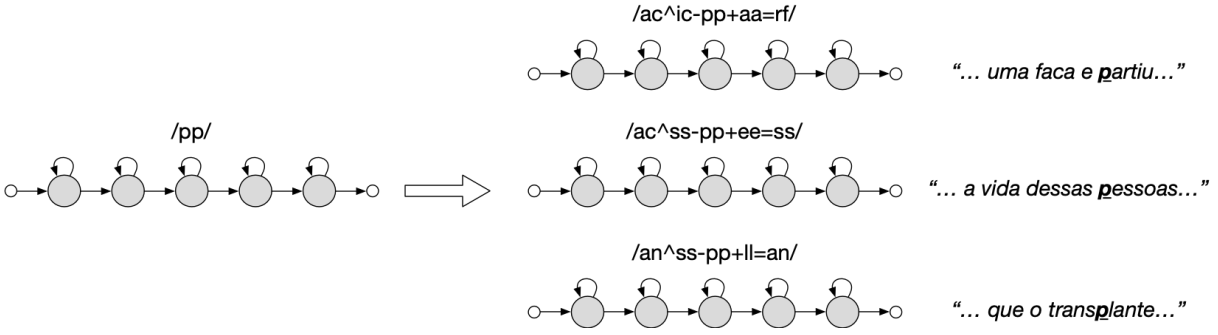


Figura 3.10 – HMMs dependentes de contexto são treinados usando como ponto de partida os modelos para os fonemas isolados correspondentes

3.3.4 Clusterização de modelos baseada em árvores de decisão

A depender do tamanho do contexto fonético utilizado, o número de fonemas dependentes de contexto pode sofrer uma explosão combinatória. Naturalmente, para um número muito alto de possibilidades, cada um dos fonemas terá poucos exemplos disponíveis na base de treinamento, o que pode comprometer a qualidade dos modelos finais.

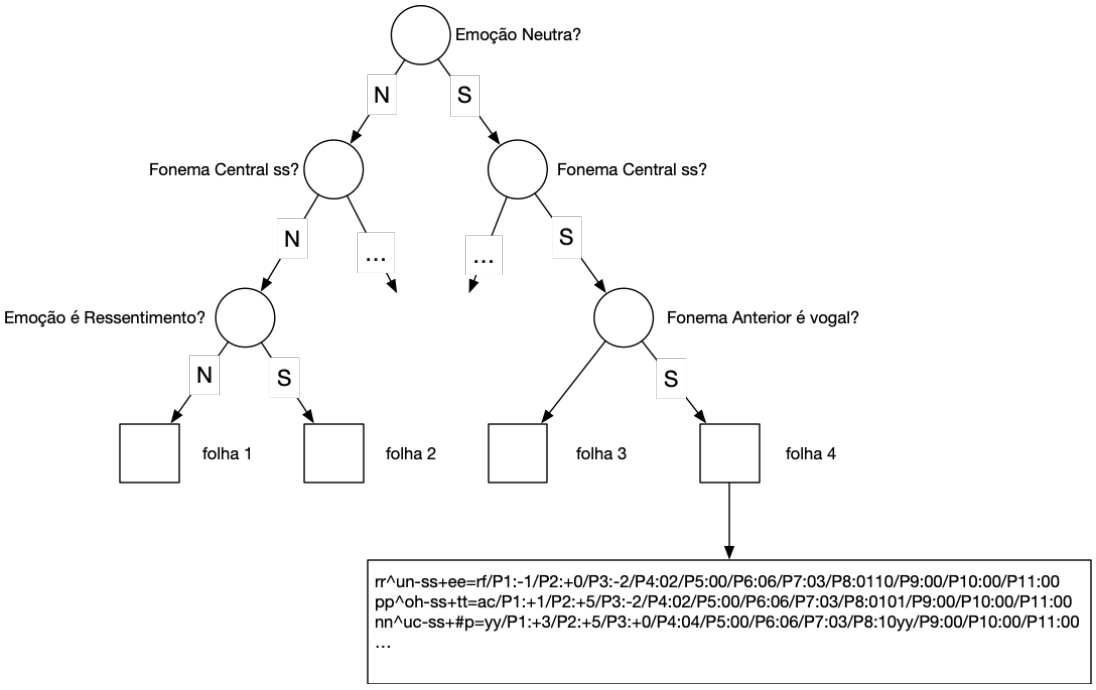


Figura 3.11 – Fonemas com vetores de atributo, de áudio e vídeo, semelhantes são agrupados de acordo com árvores de decisão. Modelos correspondentes aos fonemas agrupados numa mesma folha são treinados em conjunto

Uma alternativa para mitigar esse problema é agrupar fonemas dependentes de contexto com comportamento semelhante com o auxílio de árvores de decisão, como a mostrada na Figura 3.11. Cada nó da árvore representa uma questão relacionada ao contexto fonético. A ideia é que fonemas pertencentes a uma mesma folha terão modelos com parâmetros parecidos, e portanto podem ser treinados conjuntamente, compartilhando os exemplos de treinamento. Essa técnica foi introduzida por Young, Odell e Woodland (1994) e é utilizada na literatura de síntese de fala baseada em HMMs.

No processo de construção da árvore de decisão, todos os fonemas começam no mesmo nó. Em cada iteração, encontramos a questão que melhor separa os dados de treinamento, de acordo com alguma métrica de distância. Esta implementação utiliza a função log-verossimilhança, como descrito em (YAMAGISHI, 2006). O processo se repete, até que não haja mais questões que separem os dados de treinamento com log-verossimilhança acima de um limiar previamente definido.

Com as novas correspondências entre modelos e dados de treinamento, o reajuste

dos parâmetros dos modelos pode ser realizado mais uma vez. Devido à diferença na natureza dos atributos, a clusterização é realizada separadamente para atributos visuais e atributos acústicos. Na prática, isto significa que, para cada modelo, os exemplos de treinamento usados para ajustar a distribuição de probabilidade de emissão dos atributos visuais serão diferentes dos exemplos usados para os parâmetros acústicos.

Além de deixar o processo de treinamento mais robusto, a técnica de clusterização é importante para a etapa de síntese. Como qualquer fonema dependente de contexto sempre corresponde a alguma folha das árvores de decisão construídas, o sistema resultante será capaz de sintetizar atributos, mesmo para fonemas nunca vistos na etapa de treinamento.

3.4 Síntese

A etapa de síntese consiste em utilizar os modelos ajustados na etapa de treinamento para simular sinais temporais $\boldsymbol{F}_t$ e $\boldsymbol{V}_t$, a partir das distribuições de emissão b_i. Em teoria, isso significa que o processo descrito na Seção 3.3 pode ser utilizado para sintetizar tanto áudio quanto vídeo. Na prática, a síntese de áudio requer um volume de dados consideravelmente maior para alcançar uma qualidade satisfatória, em comparação com à síntese de vídeo. De acordo com Tokuda et al. (2013), o desenvolvimento de um sistema de síntese de voz baseado em HMMs requer algumas horas de dados de treinamento, enquanto Wang e Soong (2014) implementou um sistema de animação facial com vinte minutos de filmagens. Neste trabalho, o foco será somente a síntese da trajetória visual $\boldsymbol{V}_t$.

O processo começa com a transcrição fonética da sentença desejada, como mostrado na Figura 3.12. Os fonemas devem incluir também a representação do contexto fonético, como os fonemas usados no treinamento (ver Apêndice B). Os modelos correspondentes são arranjados de acordo com a transcrição fonética, e os sinais são amostrados de acordo com as distribuições de emissão b_i de cada estado.

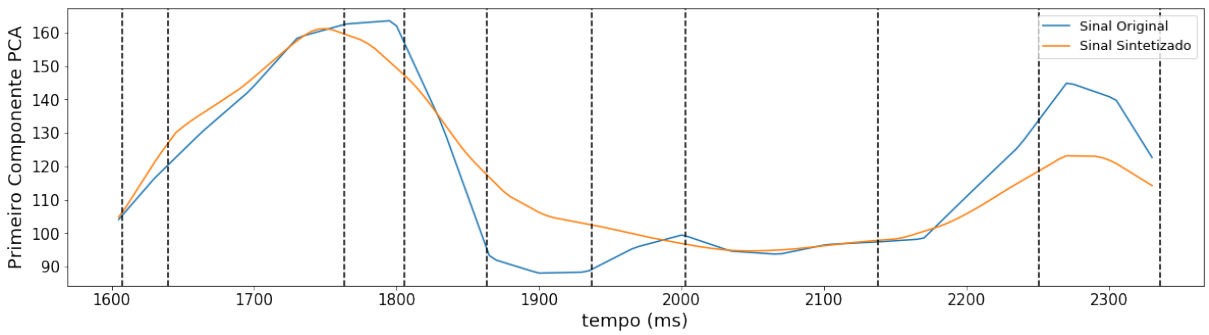


Figura 3.12 – Trajetória sintetizada a partir dos modelos treinados, em comparação com a trajetória original.

3.4.1 Modelagem de variância

Um problema recorrente com a síntese HMM é que a trajetória sintetizada pode estar excessivamente amortecida, como mostrado na Figura 3.13. Na síntese de fala, o áudio sintetizado pode ficar abafado e pouco inteligível. No caso da síntese de vídeo, os movimentos do agente sintetizado ficarão menos amplos, o que pode prejudicar a naturalidade da animação final, especialmente para a síntese expressiva, que é naturalmente mais articulada.

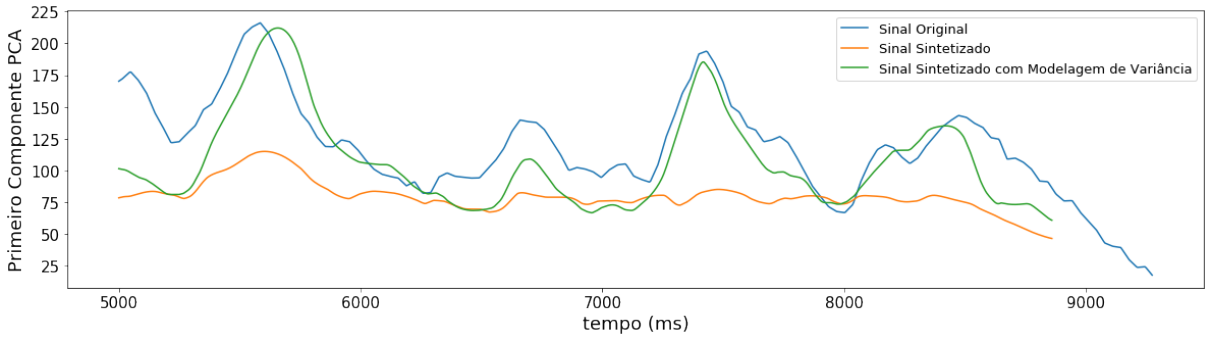


Figura 3.13 – Trajetórias sintetizadas com e sem modelagem de variância. A técnica ajuda a mitigar o problema de sobreamortecimento, comum na síntese HMM

Toda e Tokuda (2007) propõem como solução para o problema, a modelagem da variância do sinal. A ideia geral é incluir a variância do sinal como um fator de penalidade na etapa de síntese. Na prática, o processo de síntese buscará uma sequência de observações $\boldsymbol{Y}^*$ que satisfaça a Equação 3.23. O efeito no sinal final é o de preservar características de variância do sinal original.

$$\boldsymbol{Y}^* = \arg\max_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{Q}, \lambda) \cdot P(var(\boldsymbol{Y})|\lambda) \tag{3.23}$$

3.5 Considerações Finais

Este capítulo descreveu com detalhes o processo de construção de um sistema baseado em Modelos Ocultos de Markov, desde a extração de atributos da base de vídeos, até o processo de treinamento e síntese. O sistema descrito é capaz de sintetizar trajetórias de atributos visuais de animação facial expressiva.

A obtenção das animações finais necessita de uma etapa adicional, de conversão das trajetórias sintetizadas em uma sequência de imagens fotorrealistas. Esta segunda etapa não é o foco do presente trabalho, mas é importante ressaltar que o sistema, como descrito, é independente, e pode ser utilizado em combinação com algum outro que faça essa conversão entre atributos visuais e imagens, guiando as movimentações articulatórias.

O método de síntese HMM descrito aqui é inspirado nos trabalhos de Wang e Soong (2014) e Sato, Nose e Ito (2017). Ainda assim, foi necessário fazer uma série de modificações e escolhas técnicas, que certamente tem impacto no resultado final. São exemplos as escolhas de atributos audiovisuais, as definições de representação do idioma Português, como fonemas, contexto fonético e a listagem de possíves questões para construções das árvores de decisão. Todos esses são pontos que podem ser otimizados, mas esta otimização não é o objetivo desta pesquisa.

Dentre as escolhas técnicas feitas, estamos especialmente interessados em entender a validade do método proposto de modelagem expressiva e avaliar a eficácia da abordagem de modelagem de variância para reduzir o *oversmoothing* num sistema de síntese visual. Adicionalmente, há o questionamento mais geral, que é central neste trabalho: a síntese HMM é capaz de obter resultados com qualidade para o volume de dados disponível?

4 Resultados

Sistemas de síntese de animação geralmente são avaliados por meio de avaliações subjetivas perceptuais. Em avaliações deste tipo, voluntários humanos assistem e avaliam vídeos sintetizados pelo sistema, e os resultados são comparados com outros métodos de síntese. Essa é a forma mais robusta de obter resultados objetivos sobre aspectos subjetivos das imagens, como o nível de videorrealismo do agente, inteligibilidade da articulação labial e percepção da expressão de emoção.

Entretanto, neste trabalho, nos limitamos a implementar a síntese do sínal de vídeo e não chegamos à etapa de renderizar o sinal em imagens de fato. Nesse caso, um teste de percepção não seria uma avaliação apropriada e, por isso, nos limitamos a avaliar a qualidade da trajetória sintetizada, de acordo com as seguintes questões:

- O sinal sintetizado se aproxima do sinal extraído de um vídeo real, correspondente à mesma sequência de fonemas?
- O sinal sintetizado carrega informação suficiente para ser posteriormente convertido em imagens, incluindo realismo estático e articulação labial condizente com a sequência de fonemas definida?
- A modelagem de variância contribui para melhorar a qualidade do sinal sintetizado?
- A modelagem de emoção contribui para melhorar a qualidade do sinal sintetizado?
- O sinal sintetizado carrega informação de expressividade correspondente com a emoção definida?

O processo de treinamento descrito no capítulo anterior foi aplicado sobre um sub-conjunto de 80% dos vídeos da base CH-Unicamp. Foram treinadas versões do sistema com modelagem de expressão e sem modelagem de expressão e versões com e sem modelagem de variância do sinal, totalizando quatro versões do sistema. Os 20% restantes da base foram preservados para conjunto de teste. Tanto o conjunto de treinamento quanto o conjunto de teste contêm vídeos com expressividade neutra e vídeos com algum tipo de emoção. A avaliação foi conduzida sintetizando trajetórias de atributos correspondentes às sequências fonéticas dos vídeos do conjunto de teste e comparando com as trajetórias reais.

4.1 Avaliação Objetiva

Em termos objetivos, a comparação do sinal sintetizado $\hat{\boldsymbol{V}}$ com o sinal original $\boldsymbol{V}$ foi feita a partir de duas métricas principais: o erro quadrático médio, definido na Equação 4.1, e a correlação cruzada normalizada, definida na Equação 4.2.

$$EQM = \left|\hat{\boldsymbol{V}} - \boldsymbol{V}\right| \tag{4.1}$$

$$\bar{\rho} = \sum_{i=1}^{d} \frac{(\boldsymbol{V}_i - \mu_{\boldsymbol{V}_i})(\hat{\boldsymbol{V}}_i - \mu_{\hat{\boldsymbol{V}}_i})}{\sigma_{\boldsymbol{V}_i}\sigma_{\hat{\boldsymbol{V}}_i}} \tag{4.2}$$

onde $\boldsymbol{V}_i$ representa cada uma das $d = 20$ dimensões de $\boldsymbol{V}$, $\mu_{\boldsymbol{V}_i}$ e $\sigma_{\boldsymbol{V}_i}$ são, respectivamente o valor médio e desvio padrão de $\boldsymbol{V}_i$.

A Tabela 4.1 mostra os resultados finais das métricas. Em geral, a modelagem de expressão apresentou significativa melhora para ambas as métricas, tanto para os vídeos com emoção quanto para os vídeos com expressividade neutra. A modelagem de variância apresentou uma melhora na métrica de correlação, apesar de aumentar o erro quadrático médio em comparação com a trajetória original. A seguir, discutiremos implicações destes resultados com mais detalhes.

4.2 Análise Descritiva

Apesar de não haver uma etapa de renderização dos vídeos finais implementada, é possível converter o sinal sintetizado para imagens em preto e branco e fazer uma avaliação subjetiva da qualidade do processo de síntese. O que é mais importante aqui é garantir que o sinal codifica toda a informação necessária para sintetizar vídeos de alta qualidade.

A Figura 4.1 mostra exemplos de imagens sintetizadas. A saída, em geral, é fotorrealista, exceto pela região interna aos lábios, que apresenta um aspecto levemente

		Vídeos com Emoção		**Vídeos com Expressividade Neutra**	
		Síntese Simples	Síntese com Modelagem de Variância	Síntese Simples	Síntese com Modelagem de Variância
EQM ($\times 10^4$)	Sem Modelagem Expressiva	2.5	2.9	1.4	1.9
	Modelagem Expressiva	2.2	2.6	1.2	1.5
Correlação	Sem Modelagem Expressiva	0.34	0.53	0.17	0.28
	Modelagem Expressiva	0.68	0.84	0.21	0.32

Tabela 4.1 – Resultados finais. A modelagem de expressão melhora significativamente o erro quadrático médio e a correlação, para os vídeos com emoção. Alem disso, a modelagem de variância melhora a métrica de correlação em todos os cenários testados.

borrado. Tal fenômeno é comum a trabalhos que sintetizam a imagem a partir de uma projeção num espaço multidimensional simplificado das chamadas *eigenfaces*, obtidas a partir da redução de dimensionalidade proporcionada pela análise de componentes principais (COSTA; MARTINO, 2015). Como esta é uma região crítica para o realismo e inteligibilidade de um vídeo de fala, este efeito revela uma importante limitação. Possíveis estratégias para contornar este problema envolvem a adoção de abordagens híbridas, onde algoritmos de síntese de imagens fotorrealistas podem ser guiados pela trajetória sintetizada pelo sistema de conversão texto-parâmetros visuais [1]

Vídeos gerados com estes sinais apresentam sincronia labial e dinâmica de articulação. Além disso, o sinal aparenta diferenciar informações de diferentes emoções [2].

Figura 4.1 – Acima: exemplos de imagens sintetizadas com diferentes emoções. Da esquerda pra direita: expressão neutra, felicidade e raiva. Abaixo: exemplo de imagem extraída de vídeo original, para comparação.

[1] Ressalta-se que a saída do sistema proposto pode guiar animações de modelos faciais geométricos, ou 3D, nos quais o fotorrealismo é construído de maneira artificial.

[2] Exemplos de vídeos sintetizados disponíveis em <https://drive.google.com/drive/folders/1B7nqJXVpj94wxR3C7sf1SGOjlh4JARh0?usp=sharing>

4.2.1 Modelagem de Variância

O exemplo da Figura 4.2 ilustra o efeito da modelagem de variância sobre o sinal sintetizado. A síntese simples (curva laranja) produz um sinal que apresenta certa correlação com o sinal original. Mas o resultado aparenta estar excessivamente amortecido.

A modelagem de variância dá ao sinal sintetizado mais amplitude nas oscilações, tornando-as mais prõximas das de um sinal real. O efeito disto para as imagens finais é, igualmente, de dar maior amplitude às movimentações, deixando as expressões faciais melhor definidas, como mostra a Figura 4.3

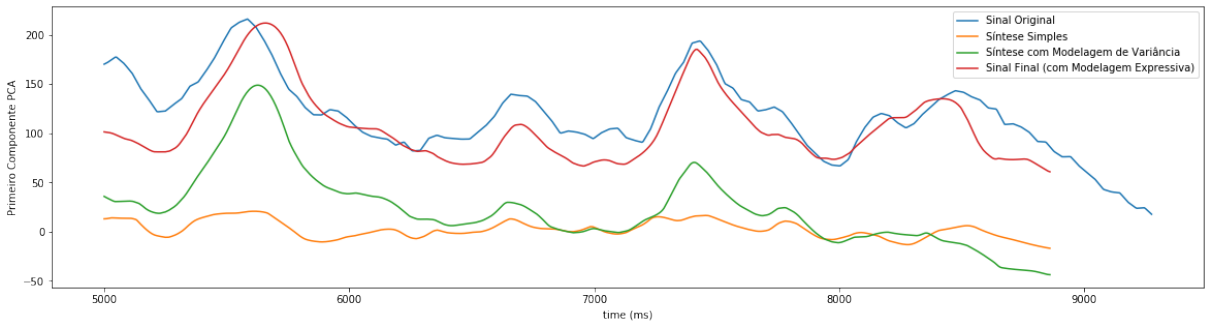


Figura 4.2 – Trajetórias sintetizadas, sob vários cenários. A modelagem de variância tem o efeito de mitigar o sobreamortecimento do sinal. A modelagem expressiva ajusta o valor médio do sinal para mais próximo do sinal original

4.2.2 Modelagem Expressiva

A Figura 4.5 mostra o resultado da interpolação de várias imagens sintetizadas, de forma a obter uma expressão facial média, para cada uma das emoções disponíveis . Como a distinção entre emoções próximas pode ser difícil, especialmente a partir de uma única imagem, uma forma alternativa de visualizar a modulação expressive é agrupar emoções visualmente similares. A avaliação de percepção de Costa (2015) encontrou, baseado nas respostas de voluntários, uma ordenação das emoções de acordo com uma valência, positiva ou negativa. A partir disso, as emoções foram classificadas em cinco grupos: fortemente negativas, negativas, neutras, positivas e fortemente positivas. A Figura 4.4 mostra as interpolações das imagens agrupadas de acordo com esta classificaçao. É possível perceber de forma mais clara, que as imagens sintetizadas seguem uma ordenação de valência, similar à observada por Costa (2015).

Figura 4.3 – A modelagem de variância tem o efeito de dar maior amplitude às movimentações e deixar as expressões faciais melhor definidas. À esquerda: quadros originais. Ao centro: quadros sintetizados sem modelagem de variância. À direita, os quadros correspondentes, mas sintetizados com a modelagem de variância.

(a) Fortemente negativas (b) Negativas (c) Neutra (d) Positivas (e) Fortemente positivas

Figura 4.4 – Expressões faciais médias, obtidas através da interpolação de várias imagens sintetizadas. As várias emoções diferentes são agrupadas por valência. (a) Fortemente negativas: raiva, nojo, censura e decepção. (b) Negativas: medo, confirmação de medos, vergonha, tristeza, pena e remorso. (d) Positivas: gratidão e amor. (e) Fortemente positivas: soberba, alívio, gratificação, admiração, esperança, satisfação, felicidade e alegria

(a) Neutro (b) Alegria (c) Felicidade (d) Esperança (e) Satisfação

(f) Alívio (g) Orgulho (h) Gratificação (i) Gratidão (j) Admiração

(k) Amor (l) Pena (m) Tristeza (n) Medo (o) Ressentimento

(p) Confirmação de medos (q) Vergonha (r) Censura (s) Remorso (t) Soberba

(u) Decepção (v) Nojo (w) Raiva

Figura 4.5 – Expressões faciais médias, obtidas através da interpolação de várias imagens sintetizadas, para cada uma das emoções.

5 Conclusão

Esta dissertação de mestrado descreve e avalia um sistema de síntese de animação facial 2D baseado em Modelos Ocultos de Markov com modelagem de expressão. As contribuições deste trabalho incluem:

- Modelagem e avaliação da expressividade para a síntese baseada em Modelos Ocultos de Markov como parte do contexto fonético;
- Construção de um sistema de síntese de animação facial baseado em modelagem estatística para o Português do Brasil. Até onde esta pesquisa pôde chegar, este é o primeiro trabalho do tipo para o idioma;
- Implementação da estratégia de modelagem de variância para a síntese de animação facial.

Uma etapa complementar de renderizar imagens videorrealistas a partir das trajetórias sintetizadas pelo sistema aqui descrito. Este trabalho futuro envolveria encontrar estratégias alternativas de representação numérica das imagens, que permitissem uma conversão direta entre vetores de atributos e vídeos de animação. Em teoria, implementar o sistema como descrito aqui, substituindo a compressão PCA por uma estratégia de representação com essas características, permitiria converter facialmente as trajetórias sintetizadas em vídeos de animação realistas.

Por opção do autor, este trabalho focou somente no aspecto da síntese de uma trajetória de atributos visuais fiel às características do contexto fonético. Isto significa que não chegamos aos videos finais de animação facial, e, consequentemente, não foi possível realizar avaliações subjetivas de percepção com voluntários humanos, como seria ideal nesse tipo de aplicação. Contudo, este trabalho avança o suficiente para mostrar que:

- A trajetória de atributos sintetizada pelo sistema tem correlação com o que seria esperado de uma trajetória real, e conserva informação suficiente para ser posteriormente convertido em imagens fotorrealistas;
- A modelagem de variância melhora a correlação da trajetória sintetizada com uma trajetória real;
- A modelagem expressiva melhora a correlação da trajetória sintetizada com uma trajetória real;

- A trajetória sintetizada carrega informação da expressividade desejada, sendo possível distinguir alguma valência nas emoções

É importante salientar que o sistema implementado aqui é um módulo completo e independente, que pode ser utilizado em combinação com outros sistemas que desempenhem o papel de conversão de imagens para representações vetoriais, e vice versa. Desta forma se revela a importância dos resultados apontados acima, visto que eles são críticos para a qualidade das animações que potencialmente serão renderizadas a partir deste sistema.

Dito isto, obter vídeos finais de animação facial e realizar avaliações de percepção certamente fortalecerá os resultados da presente dissertação. Logo, a exploração de estratégias de representação das imagens que permitam a renderização de vídeos finais se coloca como uma continuação evidente e importante deste trabalho. A estratégia de representação ideal é compacta e capaz de armazenar informação visual, estática e dinâmica, com perda mínima. Por se tratar de uma aplicação específica para imagens da articulação da fala, a representação ideal precisa ser otimizada para este tipo de imagens. Além disso, esta representação precisa possuir uma propriedade de localidade. Isto significa, para duas imagens, F_i e F_j, e suas respectivas representações, $\boldsymbol{V}_i$ e $\boldsymbol{V}_j = \boldsymbol{V}_i + \epsilon$, se fizermos $\epsilon \to 0$, então $F_i \to F_j$.

Há exemplos na literatura de trabalhos que aplicam Redes Generativas Adversariais para converter atributos geométricos de pontos da face em imagens (PHAM; WANG; PAVLOVIC, 2018). Há potencial também em explorar o espaço latente de VAEs (do inglês, Variational Autoencoders), que apresentam todas as características desejadas. (DAHMANI et al., 2019).

Referências

AN, S.; LING, Z.; DAI, L. Emotional statistical parametric speech synthesis using LSTM-RNNs. In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Kuala Lumpur: IEEE, 2017. p. 1613–1616. ISBN 978-1-5386-1542-3. Disponível em: <http://ieeexplore.ieee.org/document/8282282/>. Citado na página 27.

BAKER, J. The DRAGON system–An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 23, n. 1, p. 24–29, fev. 1975. ISSN 0096-3518. Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing. Citado na página 24.

BEVACQUA, E. et al. An expressive ECA showing complex emotions. p. 9, 2007. Citado na página 20.

BREGLER, C.; COVELL, M.; SLANEY, M. Video Rewrite: driving visual speech with audio. In: *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*. Not Known: ACM Press, 1997. p. 353–360. ISBN 978-0-89791-896-1. Disponível em: <http://portal.acm.org/citation.cfm?doid=258734.258880>. Citado na página 21.

CAO, Y. et al. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, v. 24, n. 4, p. 1283–1302, out. 2005. ISSN 0730-0301. Disponível em: <https://doi.org/10.1145/1095878.1095881>. Citado na página 20.

CHARFUELAN, M.; STEINER, I. Expressive Speech Synthesis in MARY TTS Using Audiobook Data and EmotionML. p. 5, 2013. Citado na página 27.

COOTES, T. et al. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, v. 61, n. 1, p. 38–59, 1995. Citado na página 33.

COSATTO, E.; GRAF, H. Sample-based synthesis of photo-realistic talking heads. In: *Proceedings Computer Animation '98 (Cat. No.98EX169)*. Philadelphia, PA, USA: IEEE Comput. Soc, 1998. p. 103–110. ISBN 978-0-8186-8541-5. Disponível em: <http://ieeexplore.ieee.org/document/681914/>. Citado na página 21.

COSATTO, E. et al. Lifelike talking faces for interactive services. *Proceedings of the IEEE*, v. 91, n. 9, p. 1406–1429, set. 2003. ISSN 0018-9219. Disponível em: <http://ieeexplore.ieee.org/document/1230217/>. Citado na página 16.

COSTA, P. *Animação Facial 2D Sincronizada com a Fala*. Dissertação (Mestrado), 2009. Citado na página 20.

COSTA, P. *Two-Dimensional Expressive Speech Animation*. Tese (PhD Thesis) — Universidade Estadual de Campinas, Campinas, 2015. Citado 3 vezes nas páginas 27, 30 e 49.

COSTA, P. D. P.; MARTINO, J. M. D. Image-Based Expressive Speech Animation Based on the OCC Model of Emotions. In: *Proceedings of the Facial Analysis and Animation on - FAA '15*. Vienna, Austria: ACM Press, 2015. p. 1–1. ISBN 978-1-4503-3530-0. Disponível em: <http://dl.acm.org/citation.cfm?doid=2813852.2813855>. Citado na página 48.

CUOZZO, L. G. D. et al. CNN-Based Phonetic Segmentation Refinement with a Cross-Speaker Setup. In: VILLAVICENCIO, A. et al. (Ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2018. v. 11122, p. 448–456. ISBN 978-3-319-99721-6 978-3-319-99722-3. Series Title: Lecture Notes in Computer Science. Disponível em: <http://link.springer.com/10.1007/978-3-319-99722-3_45>. Citado na página 34.

DAHMANI, S. et al. Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis. In: *Interspeech 2019*. ISCA, 2019. p. 2598–2602. Disponível em: <http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2848.html>. Citado na página 53.

DEY, P.; MADDOCK, S.; NICOLSON, R. A talking head for speech tutoring. In: *Proceedings of the ACM/SSPNET 2nd International Symposium on Facial Analysis and Animation - FAA '10*. Edinburgh, United Kingdom: ACM Press, 2010. p. 14–14. ISBN 978-1-4503-0388-0. Disponível em: <http://portal.acm.org/citation.cfm?doid=1924035.1924041>. Citado na página 16.

FAN, B. et al. Photo-real talking head with deep bidirectional LSTM. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, 2015. p. 4884–4888. ISBN 978-1-4673-6997-8. Disponível em: <http://ieeexplore.ieee.org/document/7178899/>. Citado 3 vezes nas páginas 17, 25 e 26.

FAN, Y. et al. TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks. p. 5, 2014. Citado na página 26.

GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2013. p. 6645–6649. ISSN: 2379-190X. Citado na página 26.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. Publisher: MIT Press. Disponível em: <https://doi.org/10.1162/neco.1997.9.8.1735>. Citado na página 25.

JELINEK, F.; BAHL, L.; MERCER, R. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, v. 21, n. 3, p. 250–256, maio 1975. ISSN 1557-9654. Conference Name: IEEE Transactions on Information Theory. Citado na página 24.

KAZEMI, V.; SULLIVAN, J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In: . [s.n.], 2014. p. 1867–1874. Disponível em: <https://openaccess.thecvf.com/content_cvpr_2014/html/Kazemi_One_Millisecond_Face_2014_CVPR_paper.html>. Citado na página 33.

LI, X. et al. Expressive Speech Driven Talking Avatar Synthesis with DBLSTM Using Limited Amount of Emotional Bimodal Data. In: . [s.n.], 2016. p. 1477–1481. Disponível em: <http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0364.html>. Citado na página 27.

MENDI, E.; BAYRAK, C. Text-to-Audiovisual Speech Synthesizer for Children with Learning Disabilities. *Telemedicine and e-Health*, v. 19, n. 1, p. 31–35, jan. 2013. ISSN

1530-5627, 1556-3669. Disponível em: <https://www.liebertpub.com/doi/10.1089/tmj.2011.0114>. Citado na página 16.

MORI, M.; MACDORMAN, K. F.; KAGEKI, N. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine*, v. 19, n. 2, p. 98–100, jun. 2012. ISSN 1070-9932. Citado na página 19.

MOULINES, E. et al. A real-time French text-to-speech system generating high-quality synthetic speech. In: *International Conference on Acoustics, Speech, and Signal Processing*. Albuquerque, NM, USA: IEEE, 1990. p. 309–312. Disponível em: <http://ieeexplore.ieee.org/document/115650/>. Citado na página 22.

ORTONY, A.; CLORE, G. L.; COLLINS, A. *The Cognitive Structure of Emotions*. [S.l.]: Cambridge University Press, 1990. Google-Books-ID: dA3JEEAp6TsC. ISBN 978-0-521-38664-7. Citado 4 vezes nas páginas 27, 28, 30 e 35.

PANDZIC, I. S. Facial animation framework for the web and mobile platforms. In: *Proceeding of the seventh international conference on 3D Web technology - Web3D '02*. Tempe, Arizona, USA: ACM Press, 2002. p. 27–34. ISBN 978-1-58113-468-1. Disponível em: <http://portal.acm.org/citation.cfm?doid=504502.504507>. Citado na página 16.

PANDZIC, I. S.; OSTERMANN, J.; MILLEN, D. User evaluation: Synthetic talking faces for interactive services. *The Visual Computer*, v. 15, n. 7, p. 330–340, nov. 1999. ISSN 1432-2315. Disponível em: <https://doi.org/10.1007/s003710050182>. Citado na página 19.

PHAM, H. X.; WANG, Y.; PAVLOVIC, V. Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network. *arXiv:1803.07716 [cs]*, mar. 2018. ArXiv: 1803.07716. Disponível em: <http://arxiv.org/abs/1803.07716>. Citado na página 53.

POVEY, D. et al. The Kaldi Speech Recognition Toolkit. p. 4, 2011. Citado na página 34.

SATO, K.; NOSE, T.; ITO, A. HMM-Based Photo-Realistic Talking Face Synthesis Using Facial Expression Parameter Mapping with Deep Neural Networks. *Journal of Computer and Communications*, v. 05, p. 50, jul. 2017. Disponível em: <https://www.scirp.org/journal/PaperInformation.aspx?PaperID=78666&#abstract>. Citado 2 vezes nas páginas 17 e 45.

SCOTT, K. C. et al. Synthesis of Speaker Facial Movement to Match Selected Speech Sequences. dez. 1994. Accepted: 2004-10-05T06:26:55Z. Disponível em: <https://trs.jpl.nasa.gov/handle/2014/33554>. Citado na página 21.

SHIMBA, T. et al. Talking heads synthesis from audio with deep neural networks. In: *2015 IEEE/SICE International Symposium on System Integration (SII)*. [S.l.: s.n.], 2015. p. 100–105. Citado 2 vezes nas páginas 25 e 26.

SUWAJANAKORN, S.; SEITZ, S. M.; KEMELMACHER-SHLIZERMAN, I. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics*, v. 36, n. 4, p. 1–13, jul. 2017. ISSN 07300301. Disponível em: <http://dl.acm.org/citation.cfm?doid=3072959.3073640>. Citado 3 vezes nas páginas 25, 26 e 27.

TAMURA, M. et al. Text-to-Audio-Visual Speech Synthesis Based On Parameter Generation From HMM. p. 4, 1999. Citado na página 24.

TAMURA, M. et al. Visual Speech Synthesis Based on Parameter Generation From HMM: Speech-Driven and Text-And-Speech-Driven Approaches. In: *AVSP*. [S.l.: s.n.], 1998. Citado na página 25.

TAYLOR, S. et al. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics*, v. 36, n. 4, p. 1–11, jul. 2017. ISSN 07300301. Disponível em: <http://dl.acm.org/citation.cfm?doid=3072959.3073699>. Citado na página 25.

TIWARI, V. MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 2009. Citado na página 31.

TODA, T.; TOKUDA, K. A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. *IEICE Transactions on Information and Systems*, E90D, jan. 2005. Citado na página 24.

TODA, T.; TOKUDA, K. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE - Trans. Inf. Syst*, p. 816–824, 2007. Citado na página 44.

TOKUDA, K. et al. Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE*, v. 101, n. 5, p. 1234–1252, maio 2013. ISSN 0018-9219, 1558-2256. Disponível em: <http://ieeexplore.ieee.org/document/6495700/>. Citado na página 43.

TOKUDA, K.; ZEN, H.; BLACK, A. An HMM-Based Speech Synthesis System Applied To English. In: . [S.l.: s.n.], 2002. p. 227–230. ISBN 978-0-7803-7395-2. Citado na página 24.

WANG, L. et al. Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection. set. 2010. Disponível em: <https://www.microsoft.com/en-us/research/publication/synthesizing-photo-real-talking-head-via-trajectory-guided-sample-selection/>. Citado na página 24.

WANG, L.; SOONG, F. K. HMM trajectory-guided sample selection for photo-realistic talking head. *Multimedia Tools and Applications*, p. 1–21, 2014. Citado 8 vezes nas páginas 17, 22, 25, 26, 29, 31, 43 e 45.

Wesley Mattheyses; LATACZ, L.; VERHELST, W. Auditory and Photo-realistic Audiovisual Speech Synthesis for Dutch. p. 6, 2011. Citado na página 21.

XUE, Y.; HAMADA, Y.; AKAGI, M. Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*, v. 102, p. 54–67, set. 2018. ISSN 01676393. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167639317303187>. Citado na página 28.

YAMAGISHI, J. Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis. *IEICE Transactions on Information and Systems*, E88-D, n. 3, p. 502–509, mar. 2005. ISSN 0916-8532, 1745-1361. Disponível em: <http://search.ieice.org/bin/summary.php?id=e88-d_3_502&category=D&year=2005&lang=E&abst=>. Citado na página 27.

YAMAGISHI, J. An Introduction to HMM-Based Speech Synthesis. p. 54, 2006. Citado 3 vezes nas páginas 36, 37 e 42.

YAMAMOTO, E.; NAKAMURA, S.; SHIKANO, K. Lip movement synthesis from speech based on Hidden Markov Models. *Speech Communication*, v. 26, n. 1, p. 105–115, out. 1998. ISSN 0167-6393. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167639398000545>. Citado na página 25.

YOUNG, S. The HTK Hidden Markov Model Toolkit: Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd*, v. 2, p. 2–44, jan. 1994. Citado na página 24.

YOUNG, S.; ODELL, J.; WOODLAND, P. Tree-Based State Tying for High Accuracy Modelling. In: *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. [s.n.], 1994. Disponível em: <https://www.aclweb.org/anthology/H94-1062>. Citado na página 42.

Apêndices

APÊNDICE A – Representação Fonética

	Fonema	Exemplos
Vogal oral	ac	cam**a**
	aa	f**a**to
	ee	faz**e**r
	eh	b**e**lo
	ic	quas**e**
	ii	aqu**i**
	oo	l**o**bo
	oh	t**o**ca
	uc	tud**o**
	uu	f**u**t**u**ro
Vagal nasal	an	irm**ã**
	en	**em**prego
	in	f**im**
	on	b**om**
	un	f**un**do
Aproximante (semivogal)	uw	ág**u**a
	ij	lábi**o**
	wn	pã**o**
	jn	pã**e**s
Fricativa surda	ff	**f**ato
	ss	**c**éu
	sh	**ch**apéu
	ts	**t**inta
Fricativa sonora	zz	ca**s**a
	vv	**v**erde
	zh	**j**oia
	dz	**d**ia
	rx	**r**ato
	rf	po**r**ta
Plosiva surda	pp	**p**ai
	tt	**t**enho
	kk	**c**om
Plosiva sonora	bb	**b**arco
	dd	**d**oce
	gg	**g**rande
Líquida	ll	**l**anche
	lh	traba**lh**o
	rr	ca**r**o
	rd	p**r**ato
Nasal	mm	**m**ar
	nn	**n**ada
	nh	vi**nh**o
Silêncio	#p	Silêncio de início de fim de frase.
	#c	Pausa de vírgula.
	#s	Pausa de ponto de vírgula.
	#k	Pausa de dois pontos.
	#e	Pausa de exclamação.
	#q	Pausa de interrogação.
	#a	Pausa de articulação.

Tabela A.1 – Representação fonética utilizada nessa dissertação, desenvolvida e gentilmente disponibilizada pela Fundação CPqD

APÊNDICE B – Contexto Fonético

Rótulos do pentafone	
f1	Fone antes do anterior
f2	Fone anterior
f3	Fone corrente
f4	Fone seguinte
f5	Fone depois do seguinte
Rótulos dos parâmetros adicionais de contexto	
p1	Distância do fone atual para o fone tônico presente na palavra: 0: fone tônico d: +1, ... ,+9,+y d: -1, ... ,-9,-y y: fone átono
p2	Distância do fone atual para o fone no início da palavra: 0: fone do início da palavra d: +1, ... ,+9,+y
p3	Distância do fone atual para o fone no final da palavra: 0: fone do início da palavra d: -1, ... ,-9,-y
p4	Define posição do fone atual na sentença (considera pausa como final de sentença): 00: sentença com apenas uma palavra átona 01: fone do início da até a primeira tônica 02: fone entre a primeira e a penúltima tônica 03: fone da penúltima até antes da última tônica 04: fone da última tônica até o fim da sentença
p5	Define fone presente em palavra de conteúdo ou palavra de função (baseado em lista de palavras): 00: fone presente em palavra de conteúdo 01: fone presente em palavra de função 02: fone presente em palavra de questão e sentença interrogativa Atualmente utiliza lista de palavras sem considerar contexto morfológico. No futuro, iremos utilizar o POS Tagger.

p6	Define quantidade de fones presentes na sentença em que fone atual está inserido (considera pausa como final de sentença): 00: sentença com 1 fone 01: sentença com 2 fones 02: sentença com 3 ou 4 fones 03: sentença de 5 a 8 fones 04: sentença de 9 a 16 fones 05: sentença de 17 a 32 fones 06: sentença com mais de 32 fones
p7	Define quantidade de palavras de conteúdo presentes na sentença em que o fone atual está inserido (considera pausa como final de sentença): 00: sentença com 1 palavra de conteúdo 01: sentença com 2 palavras de conteúdo 02: sentença com 3 ou 4 palavras de conteúdo 03: sentença de 5 a 8 palavras de conteúdo 04: sentença com mais de 8 palavras de conteúdo
p8	Define contexto morfológico da sentença utilizando pentagrama baseado em palavras de conteúdo e função. Morfologia da palavra onde fone atual está inserido já está contemplada no parâmetro P5. Este parâmetro define os dois contextos à esquerda e à direita (considera pausa como final de sentença):
p8.1	01: palavra antes da anterior é de função 00: palavra antes da anterior é de conteúdo
p8.2	01: palavra anterior é de função 00: palavra anterior é de conteúdo
p8.3	01: palavra seguinte é de função 00: palavra seguinte é de conteúdo
p8.4	01: palavra depois da seguinte é de função 00: palavra depois da seguinte é de conteúdo
p9	Define fone presente em uma frase afirmativa, exclamativa ou interrogativa: 00: fone presente em frase afirmativa 01: fone presente em frase interrogativa 00: fone presente em frase exclamativa 00: qualquer outra pontuação ou ausência de pontuação no final (vírgula, por exemplo) Atualmente não temos frases exclamativas, por isso, iremos colocar o mesmo valor de frases afirmativas.

p10	Define a modulação expressiva aplicada no vídeo 00: Expressividade Neutra 01: Happy For 02: Joy 03: Hope 04: Satisfaction 05: Relief 06: Pride 07: Gratification 08: Gratitude 09: Admiration 10: Love 11: Pity 12: Sadness 13: Fear 14: Ressentment 15: Fears Confirmed 16: Shame 17: Reproach 18: Remorse 19: Gloating 20: Disapointment 21: Disgust 22: Anger

Tabela B.1 – Código de representação do contexto fonético. Esta representação foi desenvolvida e disponibilizada a essa pesquisa pela Fundação CPqD.