Tecnologia de Captura de Movimento Facial Aplicada ao Estudo de Padrões Articulatórios da Fala

Mateus A. Chinelatto*, Paula D. Paro Costa†
Departamento de Eng. de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Email: *machinelatto@gmail.com, †paula@fee.unicamp.br

***Resumo*—A tecnologia de captura de movimento tem sua principal aplicação na indústria de jogos e filmes de cinema para transferir a movimentação de atores para personagens virtuais, porém ela vem sendo cada vez mais utilizada na realização de estudos sobre os movimentos do corpo humano. Neste artigo aplicamos uma metodologia baseada na captura de movimento facial, por meio de um equipamento especializado, para criar bases de dados e realizar análises visando identificar padrões e, posteriormente, um grupo de visemas que modele a fala em português do Brasil.**

***Abstract*—Motion capture technology has its main application in game and film industry to transfer the movement of real actors to virtual characters, however it has been increasingly applied in studies about the movements of the human body. In this article we apply a methodology based on facial motion capture, using a specialized equipment, to create databases and perform analysis in order to identify patterns and a group of visemes that can model Brazilian Portuguese speech.**

I. Introdução

A tecnologia de captura de movimento ou mocap (do inglês, motion capture) vem sendo amplamente utilizada na indústria de jogos e filmes de cinema para transferir a movimentação de atores para personagens virtuais [1]. Os movimentos da cabeça, incluindo a face, os olhos e a boca são especialmente importantes dentro deste contexto, pois são os principais meios da personagem expressar suas emoções e se comunicar com os outros [2]. Desta forma, ao criar a personagem a partir da captura de movimento de um ator real é possível atingir níveis maiores de realismo e expressividade.

Além das aplicações focadas na indústria do entretenimento, é crescente a utilização dessa tecnologia em estudos que envolvem a análise dos movimentos humanos, tanto do corpo como um todo, quanto de partes específicas como a face.

O estudo e análise dos movimentos da face é especialmente importante quando aplicado na melhoria dos sistemas de animação facial. Tais sistemas permitem implementar personagens virtuais personificados capazes de reproduzir o estilo de comunicação com o qual os humanos estão habituados, baseado na comunicação verbal e não-verbal. Estes dois tipos de comunicação agem de forma interdependente contribuindo para o entendimento da mensagem [3].

Para obter uma animação facial o mais realista possível, é necessário modelar corretamente a fala. Desta forma, é fundamental identificar a relação entre os movimentos da face e os sons emitidos durante a fala, particulares a uma língua. A cada som distintivo presente numa determinada língua é dado o nome de *fonema*, enquanto à sua representação visual, padrão de movimentação ou imagem relacionada a ele, é dado o nome de *visema*.

Tipicamente, a cada fonema de uma língua pode ser associado mais de um visema. Isso ocorre porque a produção acústica de determinados fonemas é visualmente semelhante entre si, não sendo facilmente distinguíveis visualmente sem o apoio sonoro. É o caso, por exemplo, dos fonemas [p,b,m]. Um outro fenômeno relevante é a coarticulação, que é a variação da articulação do fonema de acordo com o contexto fonético em que ele é realizado, como pode ser observado na realização do fonema [p] quando seguido das vogais "a" ou "u". Na pronúncia da palavra "pura", por exemplo, nota-se que no início da palavra os lábios se arredondam já durante a articulação do fonema [p], em preparação antecipatória à produção do [u] e que seu padrão articulatório será diferenciado ao da pronúncia da palavra "para".

Num estudo pioneiro para o português do Brasil, De Martino et al. empregaram técnicas de análise de movimento em vídeo e identificaram os visemas dependentes de contexto fonético para esta língua, propiciando um avanço no realismo de cabeças falantes, ou *talking heads*, para o português do Brasil [4]–[6]. O presente trabalho visa revisitar o estudo de identificação de visemas dependentes de contexto fonético do português do Brasil utilizando um sistema especializado de captura de movimento facial, que permite rastrear com precisão pontos (*landmarks*) na face de um informante. As principais contribuições deste trabalho até o momento incluem a construção de uma base já processada de dados de captura de movimento para a fala do português do Brasil e os resultados de experimentos preliminares de identificação de visemas.

II. Metodologia

A metodologia proposta, os equipamentos e aplicações software utilizados para o presente estudo estão sumarizados

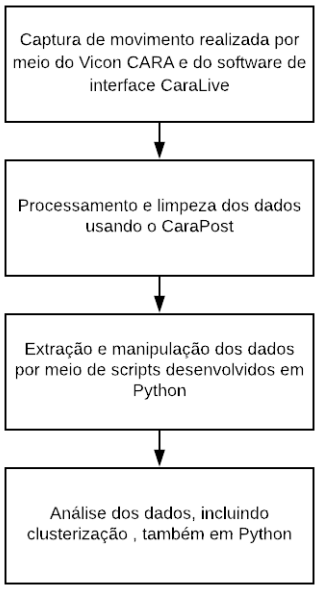

Figura 1. Fluxograma descrevendo as principais etapas e ferramentas utilizadas na metodologia proposta.

na Figura 1 e são discutidos nas seções a seguir.

A. Capturas de movimento

Para fins de comparação, foi seguida uma estratégia análoga à adotada em [4], construindo-se um *corpus* de capturas de movimento de logatomas (palavras sem sentido) paroxítonos do tipo ‘CV_1CV_2, onde C representa um som consonantal, e V_1 e V_2 representam segmentos vocálicos. Nesta configuração, segmentos consonantais são opcionais, ou seja, os logatomas também podem ser formados por duas vogais. O símbolo “ ‘ ” indica que a primeira sílaba é a tônica. Todos os fonemas referenciados neste trabalho são representados por símbolos do IPA [7]. Desta forma, foram capturados logatomas das formas:

- ‘CV_1CV_2 com C = [p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ], V_1 = [i, a, u] e V_2 = [ɪ, ɐ, ʊ], totalizando 81 casos. Dado que o tepe [ɾ] não ocorre em início de palavra, foram capturados logatomas do tipo [ɣ]V_1[ɾ]V_2 para o tratamento dos segmentos [ɾ] e [ɣ];
- ‘V_1V_2, com V_1 = [i, e, ɛ, a, ɔ, o, u] e V2 = [ɪ, ɐ, ʊ], de forma a totalizar 21 casos.

Para reduzir o número de capturas e simplificar a análise, as consoantes que compõe o corpus foram definidas explorando o chamado *ponto de articulação*. Este é o ponto do aparelho fonador em que os sons são articulados. Por exemplo, os fonemas [p, b, m] possuem o mesmo ponto de articulação, sendo produzidos na ponta dos lábios e portanto são chamados *bilabiais*. Para cada conjunto de fonemas — [p, b, m], [f, v], [t, d, n], [s, z], [l], [ʃ, ʒ], [ʎ, ŋ], [k, g], [ɾ], e [ɣ] — foi escolhido um representante. As vogais também foram escolhidas com base em critérios de semelhança e os fonemas [i, a, u] (tônicos) e [ɪ, ɐ, ʊ] (átonos) foram selecionados para ressaltar as vogais nos extremos do diagrama das vogais cardeais [4].

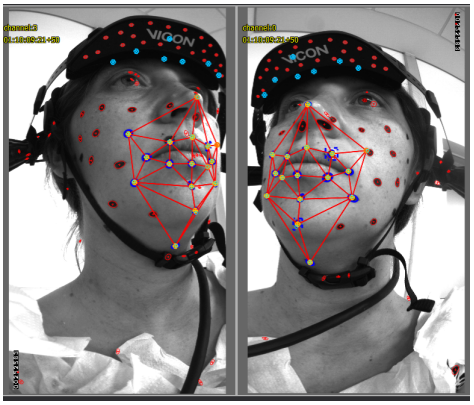

Figura 2. Imagens das câmeras inferiores durante o processamento. É possível observar os pontos rastreados para análise.

Para a realização da captura foram marcados 44 pontos na região inferior da face e foi utilizado o equipamento comercial Vicon CARA, que consiste em quatro câmeras de vídeo acopladas a um capacete e um sistema hardware/software capaz de detectar, rastrear e computar a posição espacial de marcadores na face. A captura de áudio foi realizada simultaneamente, por um equipamento independente. Posteriormente, áudio e dados de captura foram sincronizados.

B. Processamento e Análise do material capturado

Para realizar o rastreamento dos pontos desejados, montar um modelo da face, corrigir possíveis erros surgidos durante as capturas e exportar os dados coletados é necessário processar os dados no software que acompanha o equipamento, CaraPost. Por meio dele, foram selecionados 16 pontos para terem seus movimentos rastreados, dos quais se destacam os 8 localizados ao redor da boca (Figura 2). As posições tridimensionais dos pontos foram então exportadas para um arquivo do tipo *csv* (comma separated values).

Somente com os dados brutos das posições dos pontos não é possível fazer uma comparação entre os movimentos realizados em capturas diferentes. Assim, foi necessário definir uma posição inicial, e utilizá-la como referência para analisar os deslocamentos dos pontos na face durante a fala de cada fonema e observar sua trajetória.

Por fim, a partir das trajetórias dos pontos foi determinado qual frame (quadro) da captura seria considerado representante de cada som gravado.

C. Clusterização e análise dos dados

A clusterização consiste num processo de classificação de dados de maneira não-supervisionada, ou seja, sem a interferência humana nem um treinamento com dados já classificados ou rotulados, formando agrupamentos ou clusters [8].

Ela foi feita analisando os diferentes contextos de articulação de um mesmo fonema, para identificar qual ou quais os visemas que correspondem àquele som.

O método *k-means* é um algoritmo iterativo onde os clusters são determinados pela distância das amostras aos centroides dos agrupamentos. A cada iteração é calculada a posição média, ou o centro de cada um dos clusters, as amostras são então consideradas pertencentes àquele cujo centro estejam mais próximas.

III. Resultados

A. Capturas

Cada captura da forma 'CV_1CV_2 envolve dois contextos fonéticos da consoante C (CV_1 e V_1CV_2), um contexto da vogal tônica V_1 (CV_1C) e um contexto da vogal átona V_2 (CV_2). Capturas da forma 'V_1V_2 contém um contexto de cada vogal. Assim, a partir das 102 capturas realizadas estão disponíveis para análise os seguintes contextos fonéticos:

- Consoantes C = [p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ]: 12 contextos, sendo eles C[i], C[a], C[u], [i]C[ɪ], [i]C[ɐ], [i]C[ʊ], [a]C[ɪ], [a]C[ɐ], [a]C[ʊ], [u]C[ɪ], [u]C[[ɐ], e [u]C[ʊ]. Os contextos C[i], C[a], C[u] ocorrem três vezes cada e servem como parâmetro para comparar resultados futuros.
- Vogais tônicas V = [i, a, u]: também 12 contextos, sendo eles [p]V[p], [t]V[t], [k]V[k], [f]V[f], [s]V[s], [ʃ]V[ʃ], [l]V[l], [ʎ]V[ʎ], [ɣ]V[ɾ], V[ɪ], V[ɐ], e V[ʊ].
- Vogais átonas V = [ɪ, ɐ, ʊ]: 16 contextos sendo eles [p]V, [t]V, [k]V, [f]V, [s]V, [ʃ]V, [l]V, [ʎ]V, [ɣ]V], [i]V, [e]V, [ɛ]V, [a]V, [ɔ]V, [o]V, e [u]V;

B. Processamento do Material Capturado

1) Processamento e rastreamento dos pontos: Cada captura foi processada individualmente para que fossem ajustados parâmetros da detecção como o diâmetro máximo e mínimo dos pontos, o limite para a deformação do que seria considerado um marcador ou não, brilho da imagem, etc. Além disso, também foi necessário selecionar em cada câmera quais pontos deveriam ser rastreados como um único ponto e montar o modelo tridimensional da face a partir dos pontos, criando a malha de triângulos vista na Figura 2.

Para agilizar o processo, processadas as primeiras capturas, as próximas foram iniciadas a partir dos parâmetros ótimos obtidos para as anteriores.

2) Extração dos dados: Os arquivos gerados pelo CaraPost são lidos apenas por ele, desta forma foram escritos scripts para extrair os dados das posições tridimensionais dos pontos, organizá-los e salvá-los num formato mais acessível e prático para análises posteriores, no caso *csv*. Nestes arquivos os dados estão divididos em 49 colunas organizadas da seguinte forma:

- A primeira coluna corresponde ao número do quadro
- As 48 seguintes correspondem as posições nas dimensões x, y e z de cada um dos 16 pontos, organizadas sempre na ordem xyz, por exemplo: ponto1x, ponto1y, ponto1z, ponto2x, ponto2y, ponto2z...

Os scripts utilizam uma API chamada PyCara e a biblioteca *open source* de análise de dados Pandas para ler o arquivo num *DataFrame*. Os arquivos resultantes desta fase constituem uma base de dados de capturas processadas, porém ainda apresentam apenas os dados brutos capturados pelo equipamento como pode ser visto na Figura 3.

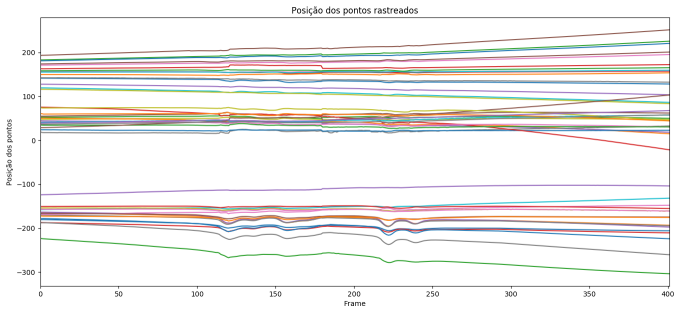

Figura 3. Trajetórias dos pontos rastreados durante a fala da frase: "Paula fala papa". Cada curva representa o deslocamento de uma coordenada de um determinado ponto. Logo existem 48 curvas contendo o movimento em x, y e z de cada um dos 16 pontos. É possível distinguir bem três grupos que correspondem a x, y ou z.

A pose de referência foi definida como um dos pontos de silêncio, ou posição de repouso, logo antes do início da articulação do logatoma. Ela foi definida individualmente para cada uma das 102 capturas, e foi tomada durante o silêncio entre a frase de controle "Paula fala" e o logatoma. Por exemplo, na captura representada na Figura 3 o frame escolhido foi o número 205. A figura 4 apresenta o deslocamento dos pontos marcados na face durante a articulação da palavra "papa" a partir da posição definida como repouso.

3) Determinação dos Alvos Articulatórios: A caracterização fonética dos segmentos é tradicionalmente efetuada através da descrição de alvos articulatórios [9]. Um alvo articulatório pode ser entendido como uma representação estática da conformação do trato vocal característica de um segmento. Desta maneira cada fonema foi associado ao frame correspondente ao seu alvo articulatório.

A determinação do quadro correspondente foi feita analisando qual o ponto e qual sua dimensão (x, y ou z) de maior deslocamento. A partir desta dimensão os alvos articulatórios foram então definidos como os frames correspondentes aos picos de deslocamento. Identificados os alvos articulatórios, eles foram relacionados aos fonemas correspondentes e ao contexto fonético que representam. Os dados sobre a posição tridimensional de cada ponto nos frames identificados foram agrupados por fonema, ou seja, foi criada uma nova base de dados organizada da seguinte maneira:

- Cada arquivo corresponde aos contextos referentes a um fonema, por exemplo o arquivo do fonema "p" possui as informações de 18 frames representando alvos articulatórios, sendo C[i], C[a], C[u] repetidos três vezes totalizando assim os 12 contextos citados anteriormente.
- As informações estão dispostas em 51 colunas, das quais a primeira indica o contexto fonético, a segunda a captura de qual se origina o contexto, a terceira o frame correspondente na captura e as outras 48 às posições tridimensionais dos pontos.

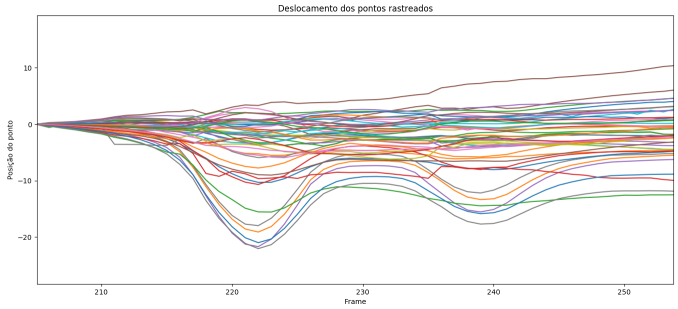

Figura 4. Deslocamento dos pontos durante a fala da palavra "papa", tomando como referência para o repouso o frame 205. Novamente cada traço representa o deslocamento de uma coordenada de um determinado ponto.

A clusterização foi utilizada para identificar diferentes padrões articulatórios de um mesmo fonema. Ela revela quais contextos possuem articulações visualmente parecidas e quais são distinguíveis.

O método escolhido foi o *k-means*, ele foi aplicado considerando todos os 48 valores (coordenadas x, y e z de cada um dos 16 pontos). Inicialmente, o número k de agrupamentos (clusters) foi definido aplicando-se o método do cotovelo (*elbow method*), que consiste em executar o algoritmo de clusterização sucessivas vezes, aumentando o número de clusters a cada iteração, computando-se uma métrica que indique se houve melhoria na discriminação dos dados com o aumento do número de clusters. Neste trabalho, adotou-se como métrica a soma das distâncias ao quadrado de cada ponto ao centróide do cluster mais próximo.

4) Análise da clusterização do fonema [f]: Ao utilizar o *elbow method* para determinar o número ótimo de grupos do fonema [f] esse número ficou entre 4 e 5 clusters.

Como citado anteriormente os contextos [fa], [fi] e [fu] foram realizados três vezes. Esperava-se que os contextos que fossem iguais entre si estivessem num mesmo cluster, porém ao utilizar quatro ou mais clusters isso não ocorre. Este fato levou a uma análise com três, dois e um único grupo.

- Com três grupos o contexto [afɐ] aparece sozinho em um cluster, assim como o contexto [afɪ], todos os outros [fi] [fa] [ifɪ] [ifɐ] [ifʊ] [fu] [afʊ] [ufɪ] [ufɐ] [ufʊ] se encontram no mesmo grupo.
- Com dois grupos apenas o contexto [afɪ] aparece num cluster diferente.

Ao observar diretamente os dados as capturas [fafɐ] e [fafɪ] eles realmente se destacam dos outros. Caso eles sejam desconsiderados, todos os contextos fonéticos do fonema [f] se encaixam no mesmo grupo.

Esse resultado, bem como o encontrado para quatro clusters se distancia do esperado, dessa forma surge a necessidade de analisar novamente os dados e os resultados das clusterizações considerando novas hipóteses e eliminando possíveis erros. Outro ponto levantando durante esta análise é que a posição definida como repouso pode ter influenciado diretamente nos resultados obtidos, sendo assim seria interessante realizar uma nova análise escolhendo outra forma de definir a posição de silêncio e comparar os resultados para buscar incoerências.

IV. Conclusão

Ao todo foram analisadas 102 capturas, processo que resultou na criação de uma base de dados com capturas já processadas bem como um conjunto de scripts em Python para extrair e reorganizar os dados referentes às posições tridimensionais dos pontos marcados na face. Tais capturas e scripts constituem um dos resultados deste trabalho pois os mesmos podem ser reutilizados em estudos semelhantes.

Os alvos articulatórios foram identificados e os dados referentes a eles foram compilados e utilizados para identificar quais parões articulatórios de um mesmo fonema são visualmente distintos e portanto representados por visemas diferentes. Os resultados apresentados não devem ser tratados ainda como definitivos. A definição da posição de silêncio pode ter influenciado diretamente nos resultados.

A perspectiva é dar continuidade ao trabalho, levantando novas hipóteses sobre os resultados já obtidos e realizando a clusterização para todos os fonemas capturados.

Agradecimentos

Os autores agradecem Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro dado a este projeto e ao Galileu, Laboratório de Computação de Alto Desempenho e Ambiente 3D de Visualização Científica da Unicamp.

Referências

[1] A. Menache, *Understanding motion capture for computer animation and video games*. Morgan kaufmann, 2000.

[2] K. Ruhland, M. Prasad, and R. McDonnell, "Data-Driven Approach to Synthesizing Facial Animation Using Motion Capture," *IEEE Computer Graphics and Applications*, vol. 37, no. 4, pp. 30–41, 2017.

[3] J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan, and N. Badler, "Creating interactive virtual humans: Some assembly required," *Intelligent Systems, IEEE*, vol. 17, no. 4, pp. 54–63, 2002.

[4] J. M. De Martino, L. P. Magalhães, and F. Violaro, "Facial animation based on context-dependent visemes," *Computer & Graphics*, vol. 30, pp. 971–980, 2006.

[5] P. D. P. Costa and J. M. D. Martino, "Assessing the Visual Speech Perception of Sampled-Based Talking Heads," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[6] P. Costa, "Two-dimensional expressive speech animation," Ph.D. dissertation, PhD thesis, School of Electrical and Computer Engineering, 2015.

[7] D. M. Decker *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000.

[9] T. Association, *Handbook of the international phonetic association-a guide to the use of the international phonetic alphabet.* Cambrige University Press, 1999.