Protótipo de Correlação de Eventos de Leitura Labial com Medidas Temporais e Legendas

João Marcelo S. Souza [**,*] **Caroline da Silva M. Alves** [**,*]
Orlando Mota Pires [*] **Jés de Jesus F. Cerqueira** [**]
Wagner Luiz A. de Oliveira [**]

[*] *Centro de Supercomputação SENAI CIMATEC, Salvador, Bahia, Brasil (e-mails: joao.marcelo@fieb.org.br, joaomssouza@gmail.com, sma.caroline@ymail.com, orlandomota2002@gmail.com*
[**] *Programa de Pós-Graduação em Engenharia Elétrica, UFBA, Salvador, Bahia, Brasil (e-mails: jes@ufba.br, oliveira.wagner@ufba.br)*

Abstract: Computer vision has become a prominent scientific field in recent years due to the various applications enabled by artificial neural networks. However, to make these networks feasible, having qualified and labeled data is essential, which is often more complex for some application possibilities, especially those that evaluate the temporal evolution of movements. Therefore, this research proposes a prototype methodology to correlate the temporal effects of mouth movements obtained by visually extracted measures via video streaming with multimodally labeled speech and lip-reading events from video subtitles. With the universe of data extracted from public sources, indications were observed that it is possible to represent a repeated word over the same video via time series. The experiments brought new challenges regarding the synchronization quality between video images and the respective start and end times of words in subtitles. Also, the same word may be pronounced with different durations, making comparing the samples more difficult. The results enable future development of datasets of temporal series representing words visually through standardized measures and using captions from diverse videos to increase the range of sample classes. Consequently, such aspects contribute to creating applications using time series training.

Resumo: Visão computacional se tornou uma área científica de evidência nos últimos anos em função das diversas aplicações disponibilizadas principalmente pelas redes neurais artificiais. Contudo, para viabilizar estas redes é primordial ter dados qualificados e rotulados, fato que não é consenso para todas as possibilidades de aplicações, em especial aquelas que avaliam a evolução temporal de movimentos. Neste sentido, é proposto um protótipo de metodologia que busca correlacionar os efeitos temporais de movimentação de boca, obtidos por medidas extraídas de forma visual via vídeos, com eventos de leitura labial rotulados de forma multimodal com as legendas dos vídeos. Com o universo de dados extraídos de fontes públicas, foram observados indícios de que é possível representar via séries temporais uma palavra repetida ao longo do mesmo vídeo. Os experimentos trouxeram novos desafios quanto à qualidade de sincronização entre as imagens dos vídeos e tempos de início e fim das palavras nas legendas e também, foram encontradas durações diferentes para a representação da mesma palavra, o que dificulta as comparações diretas das amostras. Os resultados possibilitam desenvolvimentos futuros de *datasets* de séries temporais, que representam palavras de forma visual (através de medidas padronizadas) e usam legendas de vídeos diversos para aumentar a capilaridade das classes de amostras, e consequentemente, contribuem para criação de aplicações com uso de treinamento de séries temporais.

Keywords: lip reading; visual extraction; standardized measurements; subtitles; labeling.
Palavras-chaves: leitura labial; extração visual; medidas padronizadas; legendas; rotulagem.

1. INTRODUÇÃO

Nos últimos anos, o aperfeiçoamento da resolução e da taxa de quadros por segundo das câmeras, aliados aos algoritmos de aprendizagem de máquina e inteligência artificial, têm tornado os sistemas de visão computacional cada vez mais fidedignos, viabilizando o uso de câmeras como uma fonte alternativa e poderosa, sem contato e não-invasiva, para aquisição e estimação de sinais biológicos de humanos (Noceti et al., 2020).

Neste cenário, surgiram diferentes aplicações envolvendo detecção de padrões como expressões faciais (Dalvi et al., 2021), comportamentos de movimentação humana (Chasmai et al., 2022) e leituras bio-cinéticas como a labial (Chung e Zisserman, 2018; Fenghour et al., 2021). Neste contexto, a leitura labial de forma automatizada tem alta relevância no contexto da triagem de sinais humanos, com respectivas aplicações nas áreas de saúde e psicologia, uma vez que é possível inferir componentes de fala em ambientes com alto ruído e com baixa captação de sons.

No entanto, para viabilizar este sistema computacional de mensuração de sinais humanos é necessário mitigar alguns efeitos temporais causados pelas movimentações entre câmera e face humana, e também reduzir as condições adversas de distorções, perspectivas e oclusões ocasionados em função da relação de aquisição bidimensional (2D) em um ambiente real tridimensional (3D) (Noceti et al., 2020; Dalvi et al., 2021; Chasmai et al., 2022).

Com base nestes desafios, este trabalho propõe investigar se é possível identificar, estabilizar pontos de interesse de órgãos humanos, realizar a mensuração destes ao longo do tempo e, ao final, correlacionar os sinais das séries temporais das medidas com eventos temporais de leitura labial estimados a partir da extração de legendas de vídeos de fontes públicas. Esta é uma abordagem pioneira visto que trabalhos anteriores fazem uso de legendas como fonte para rotulagem de eventos de imagens para entrada de redes neurais, como em Chung e Zisserman (2018), ao invés das medidas relativas dos órgãos ao longo do tempo, como proposto neste trabalho. Diante da carência de dados qualificados e de referências rotuladas de sinais temporais humanos, a proposta deste trabalho também corrobora para reforçar o uso de legendas de vídeos como alternativa auxiliar de rotulagem e correlação de eventos temporais de fala.

O trabalho é composto por dois grandes blocos funcionais. O bloco 1 é responsável pelo pré-processamento dos dados: extração, estabilização e normalização de pontos de interesse correspondentes à face, seguidos da estimação de pontos de contorno da boca e suas respectivas medidas, que contribuem para a caracterização dos movimentos durante a fala. O bloco contempla sistemas que têm como entrada vídeos de fontes públicas e, como saída, um conjunto de medidas da boca ao longo do tempo, gerando um dataset de séries temporais. Já o bloco 2 é responsável pela extração de palavras nas legendas dos vídeos de fontes públicas, determinando os respectivos tempos de início e fim de cada palavra nas sentenças do vídeo. Em resumo, a partir desta abordagem é possível verificar se os sinais medidos no bloco 1 podem servir de base para representar e inferir o comportamento ao longo do tempo das palavras extraídas das legendas no bloco 2.

Na Seção 2 são apresentados os trabalhos correlatos, na seção 3 é apresentada a proposta, os resultados são encontrados na Seção 4, seguido das conclusões, Seção 5.

2. TRABALHOS RELACIONADOS

A aplicação de câmeras digitais e sistemas inteligentes na estimativa de expressões faciais tem sido amplamente explorada na comunidade científica (Dalvi et al., 2021; Happy e Routray, 2015). Essas aplicações frequentemente seguem princípios básicos de rastreamento facial e detecção de expressões, como abordado em Dalvi et al. (2021), onde avalia-se o efeito das emoções de forma estática em cada quadro de um vídeo separadamente, ou seja, cada quadro tem sua respectiva detecção. Os principais desafios dessas abordagens estão na conexão temporal entre cada quadro de uma sequência e na escassez de *datasets* que consideram os efeitos temporais, como também na obtenção de rótulos precisos e validados, fatos que dificultam o uso em estudos e treinamentos de redes neurais.

Na análise visual do discurso os desafios estão relacionados à identificação do orador em um diálogo com vários sujeitos e a importância da característica temporal para a caracterização da dinâmica do movimento (Zhou et al., 2014). Outros desafios estão relacionados à variabilidade das expressões faciais durante a fala e à ambiguidade visual resultante de movimentos idênticos para a pronúncia de diferentes palavras (Sheng et al., 2024).

O processo de extração de biosinais das deformações dos músculos faciais começa com a identificação das regiões de interesse (ROIs). Um método comum para essa identificação é o uso de pontos de interesse (do inglês, *landmarks*) (Fenghour et al., 2021). A biblioteca Dlib (King, 2009) disponibiliza um modelo de predição de *landmarks* baseado no método Kazemi e Sullivan (2014), que é amplamente utilizado para identificar a forma do rosto.

O segundo passo envolve a extração de características relevantes para o objetivo da análise. Segundo Dupont e Luettin (2000), na análise visual do discurso, uma abordagem comum é a extração geométrica de características. Essa abordagem se concentra em medidas entre partes da boca, como largura e altura, para a representação dos movimentos, como em Kaynak et al. (2004).

A identificação dos movimentos labiais pode ser feita em classe de caracteres ASCII, visemas ou palavras (Sheng et al., 2024). O uso de caracteres ASCII ou palavras como classes pode resultar em previsões limitadas, pois o sistema só pode reconhecer as letras e palavras que foram treinadas. Por outro lado, o uso de visemas como classes oferece uma abordagem mais flexível, permitindo a generalização para palavras não vistas durante o treinamento e para diferentes idiomas. Visemas são equivalentes visuais de fonemas (segmentos de som na fala) e, geralmente, são menos numerosos que o conjunto de caracteres ASCII. No entanto, a classificação de visemas apresenta desafios devido à sua curta duração e ao fato de um visema representar mais de um fonema, resultando em ambiguidade visual, o que requer um grande volume de dados de treinamento conforme destacado por Fenghour et al. (2021).

3. PROPOSTA

Este trabalho propõe investigar a hipótese de que é possível relacionar a medição visual de pontos de interesse humanos ao longo do tempo com padrões de fala, viabilizando a construção futura de sistemas de leitura labial. Para tanto, foi montado um ambiente de prototipação e testes que tem as seguintes características: (I) realizar a extração, estabilização e normalização espaço-temporal de pontos da boca humana capturados ao longo do tempo através de *streaming* de vídeo de fontes públicas para, ao final, fornecer séries temporais multivariadas referentes às medidas da boca; (II) capturar e armazenar as legendas contendo as frases/sentenças, bem como os respectivos instantes de início e fim, sincronizados de acordo com a linha do tempo dos vídeos.

Com este ambiente é possível estimar, por exemplo, qual é o comportamento da movimentação da boca a partir da observação de uma determinada palavra previamente rotulada pela legenda. Esta estimação se torna factível em função da busca de similaridades entre os sinais das séries marcados ao longo do tempo e os instantes de início e fim das sentenças nas legendas.

Outro objetivo deste projeto é desenvolver ferramentas para mitigar os efeitos da alta variabilidade ocasionada pelas condições adversas de movimentações, das limitações oriundas da relação de amostragem 2D da câmera no espaço real 3D, tais como rotações e efeitos de perspectivas. O foco é gerar sinais padronizados, garantindo critérios comparáveis, mensuráveis e repetíveis que possam contribuir para uma análise analítica das medidas da boca e face humana.

A Figura 1 apresenta a arquitetura do sistema computacional esperada a partir dos conceitos propostos neste trabalho, que, em geral, contempla os seguintes passos enumerados de acordo com a respectiva marcação na figura: (A) remoção da imagem de fundo, identificação e focalização da face; (B) extração de pontos de interesse (*landmarks*) que caracterizam os principais órgãos da face, como boca, olhos e nariz; (C) normalização, que corresponde a estimação de descritores espaço-temporais, calibração, correção e marcações relacionadas às perspectivas e rotações, entregando ao final sinais em um formato estável; (D) geração das medidas, que caracterizam movimentos em amplitudes ao longo do tempo, definidas por meio de séries temporais; (E) inferência que estime, com base no padrão das séries, uma palavra ou conjunto delas.

Para gerar dados de validação e testes foi utilizada uma metodologia que contempla os seguintes passos, enumerados de acordo com a Figura 2:

I. Definir conjunto de vídeos de fontes públicas, onde são extraídas:
   a. As séries temporais que correspondem a medidas da boca, de acordo com os passos de (A) a (D) da Figura 1, gerando as "*séries temporais brutas*";
   b. As legendas de todos os vídeos, com a respectiva marcação temporal de início e fim para sentença/frase.
II. Gerar um histograma das sentenças ao calcular a frequência das palavras mais utilizadas com base em um *data lake* simplificado obtido a partir da varredura da legenda de todos os vídeos;

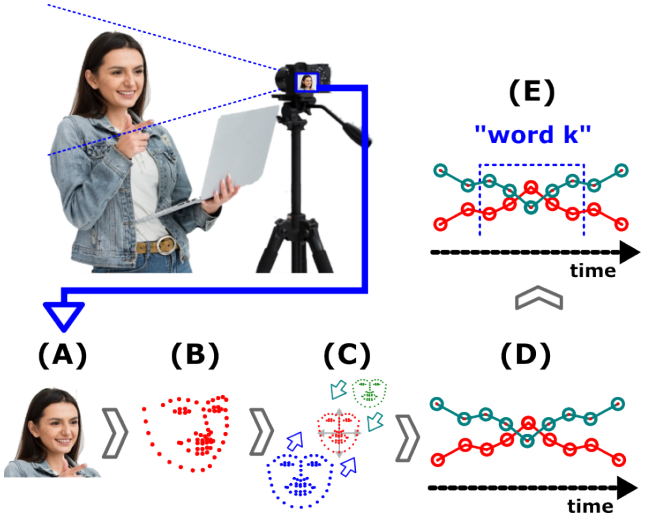

Figura 1. Diagrama funcional do conceito da arquitetura que relaciona as medidas características de movimento, extraídas de forma visual, em formato de séries temporais com a identificação da subsérie que representa a assinatura de uma palavra de interesse (imagem dos autores, adaptado de Freepik.com).

III. Definir a partir de análise manual em amostras dos vídeos ao longo do tempo, para prototipagem, qual palavra apresenta mais características evidentes de movimentação, a qual será denominada *"palavra semente"*;
IV. Identificar as ocorrências da *"palavra semente"* nos arquivos de legendas dos vídeos, com a indicação dos tempos de início e fim das sentenças/frases que a contém;
V. Com as ocorrências e tempos de início/fim das sentenças que contém a *"palavra semente"*, realizar o recorte destes instantes nas séries temporais de todos os vídeos gerados a partir do passo I.a, gerando assim as *"séries temporais recortadas"*;
VI. De forma manual, extrair as séries temporais das medidas *"palavra semente"*, com base nos resultados do passo V, gerando as *"séries de referência semente"*;
VII. Realizar análise estatística que considera os seguintes aspectos:
   a. Observar se as *"séries de referência semente"* estão contidas no universo das *"séries temporais recortadas"*, originadas dos recortes das legendas, a partir da busca cruzada de similaridade de padrões no mesmo vídeo e nos demais vídeos. Como padrão, o início e fim corresponde à legenda da sentença/frase e não diretamente à palavra, desta forma, será confrontado se sinais da palavra estão contidos nos sinais da frase;
   b. Avaliar se as *"séries de referência semente"* estão contidas no universo das *"séries temporais brutas"* oriundas dos vídeos inteiros transcritos em séries temporais.

A partir destes passos da metodologia também é possível avaliar os graus de similaridade das palavras faladas de acordo com o comportamento comparativo das suas respectivas séries. Deste modo, tal abordagem poderá via-

bilizar posterior utilização deste conjunto como entrada para o treinamento de redes neurais que estimem palavras faladas em função das séries temporais.

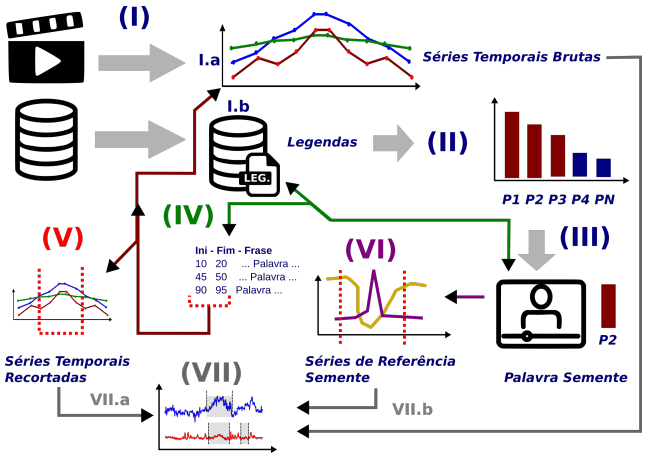

Figura 2. Fluxograma da metodologia de correlação de séries temporais a partir de legendas (imagem dos autores).

4. RESULTADOS

4.1 Medidas Básicas

Para viabilizar a validação preliminar das hipóteses, primeiramente foi desenvolvida uma metodologia simples para realizar a estimação das movimentações da boca a partir dos 68 pontos de interesse extraídos através da ferramenta Dlib (King, 2009). Os resultados deste processo estão apresentados na Figura 3.

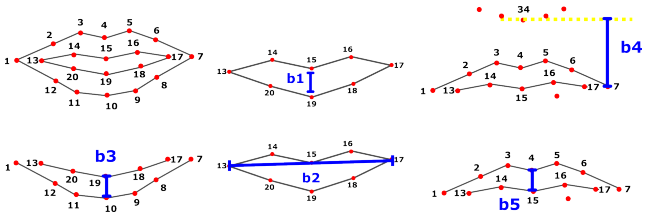

Figura 3. Exemplo de pontos que representam a boca com a indicação de algumas medidas (de b1 a b5) correspondentes à abertura interna vertical e horizontal, aos lábios superiores e inferiores, e à relação da boca com os pontos da face, em específico do nariz.

Inicialmente, foram escolhidos os pontos de interesse que representam a boca e pontos de referência com relação à face. Por exemplo, o ponto b4 da Figura 3 utiliza um ponto do centro do nariz como referência, pois é uma área com pouca variação muscular, conforme indicado nas Unidades de Ação associadas ao sistema FACS (do inglês, *Facial Action Coding System*), desenvolvido inicialmente por Ekman et al. (2002).

Uma vez definidas as medidas básicas, foi realizado o processo de normalização, de acordo com pontos apresentados em Dalvi et al. (2021), de modo a estabilizar os efeitos de movimentação ao longo da cena. Após o desenvolvimento destes módulos, foram extraídos vídeos de fontes públicas com suas respectivas legendas, que sumarizaram cerca de 5 mil segundos e 1400 palavras, em um montante de 30 vídeos do Youtube. No geral estes vídeos têm as seguintes características: um único sujeito em destaque na cena, foco em conversação, sem muitas transições e efeitos visuais, e todos são em língua inglesa. Para o processo de extração quadro a quadro de todos os vídeos foi utilizado cerca de 10 horas de 1 (um) nó computacional de CPU de 40 núcleos e 196 GB de RAM.

4.2 Sincronização de Legendas e Sinais

A Figura 4 mostra as palavras mais recorrentes encontradas nas legendas dos vídeos analisados.

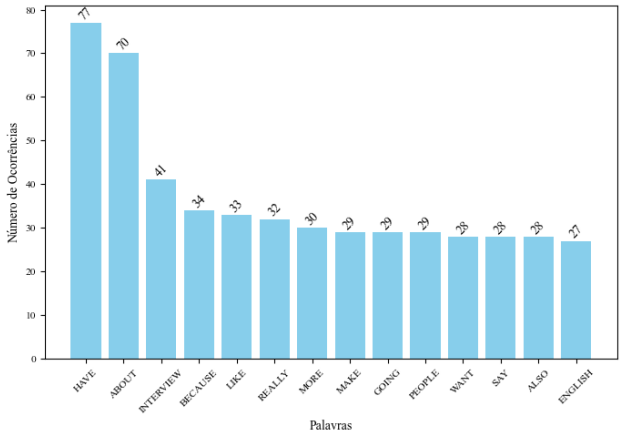

Figura 4. Histograma com a frequência de ocorrências das 15 palavras mais utilizadas.

Dentro do conjunto de palavras mais recorrentes encontradas, foram agrupadas as palavras auxiliares, classificadas como "*stop words*", e as com maior frequência. O termo "*Stop words*" se refere a palavras comuns que não contribuem significativamente para o significado do texto (Ladani e Desai, 2020). Essas palavras correspondem a 1173 ocorrências e estão presentes em 100% dos vídeos.

Após a remoção das "*stop words*", foram selecionadas para análise as três palavras que apresentaram maior frequência de ocorrência. Na Figura 5, são apresentadas as assinaturas de uma amostra para cada palavra. Essas assinaturas são geradas a partir das medidas propostas (b1 a b5), que estão representadas na Figura 3.

Após uma análise visual do efeito do comportamento de cada uma das séries, como ilustrado na Figura 5, foi verificado de forma manual que a palavra "*interview*" contém a assinatura mais evidente entre as ocorrências das três palavras, sendo utilizada como protótipo para as demais etapas deste trabalho.

4.3 Busca e Correlações de Eventos

As análises das sessões anteriores correspondem aos passos I, II, III e VI, enumerados conforme Figura 2. A partir da palavra selecionada "*interview*" foram buscadas as 44 ocorrências no universo de legendas, realizando assim o passo IV. Com base nos tempos de início e fim de cada sentença/frase, foram cortadas as séries identificadas no passo I, concluindo assim o passo V, conforme Figura 2.

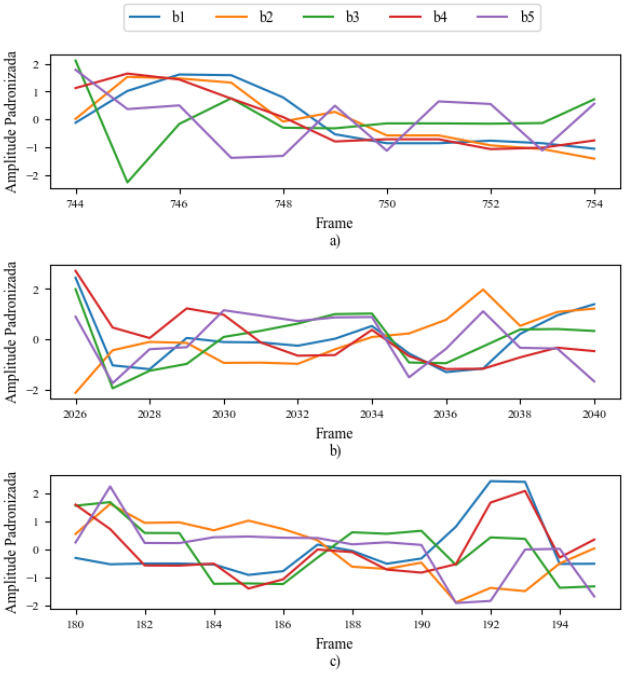

Figura 5. Séries temporais das medidas para as três palavras com maior frequência nos vídeos analisados: a) "*have*", b) "*about*" e c) "*interview*".

De posse das *"séries de referência semente"* e das *"séries temporais recortadas"* foi possível encontrar em um mesmo vídeo as variações de representação da mesma palavra *"interview"*, conforme ilustrado na Figura 6, executando assim o passo VII.a da Figura 2.

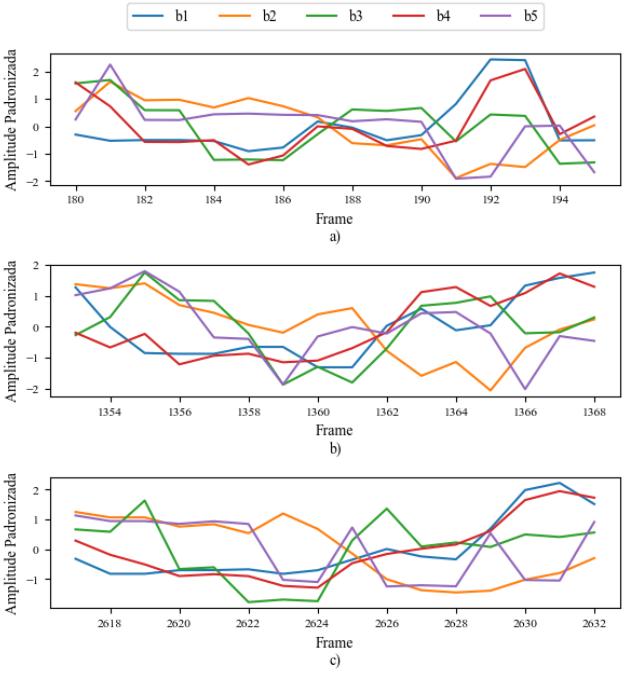

Figura 6. Sinais das medidas correspondentes às ocorrências da palavra "*interview*" encontradas no mesmo vídeo. As amplitudes foram padronizadas utilizando a normalização Z-score.

Para análise, foi utilizada a série (a) da Figura 6, como padrão de busca automatizada através a biblioteca Stumpy (Law, 2019), sendo as séries (b) e (c), da mesma classe que (a) econtradas pelo Stumpy. Como critério de seleção de semelhança, foi configurado o ponto de corte correspondente a 1,5 desvios padrão abaixo da média das distâncias euclidianas calculadas na série temporal de busca. Com o objetivo de avaliar os resultados, utilizou-se os valores de erro percentual absoluto médio (MAPE) para cada medida da boca, entre a série (a) e cada uma das séries encontradas, esses resultados são apresentados na Tabela 1.

Os valores apresentados evidenciam a imprecisão ponto a ponto no padrão entre a série semente e as séries encontradas, sendo essa diferença maior para as medidas b4 e b5. Isso se deve às variações nos movimentos dos lábios, ao movimento da cabeça e aos ruídos na inferência dos landmarks. Apesar dessas questões, foi possível detectar séries de mesma classe utilizando a distância euclidiana devido às semelhanças na evolução do movimento ao longo do tempo.

Tabela 1. MAPE entre a série semente (a) e as séries (b) e (c) apresentadas na Figura 6

	MAPE (%)	
Medidas	**(a) e (b)**	**(a) e (c)**
b1	3,95	1,72
b2	1,03	1,01
b3	1,53	1,23
b4	12,74	14,18
b5	25,08	15,92

É importante destacar que foram encontradas variações não somente no padrão de forma de cada uma das medidas, mas também na duração das ocorrências das palavras. Dessa forma, no passo VII.a da Figura 2, ao utilizar a busca automática de padrões foi possível encontrar ocorrências equivalentes apenas nas sentenças pertencentes ao mesmo vídeo da semente. Pois, as variações se intensificam nas amostras com interlocutores diferentes. Do mesmo modo que a busca de sementes no universo de todas as séries do conjunto de vídeos, passo VII.b indicado na Figura 2, somente retornou as amostras do próprio vídeo. Tais resultados a nível de protótipo foram considerados satisfatórios, contudo para generalizar a metodologia faz-se necessário investigar métodos mais robustos para busca de padrões em séries temporais.

Outro aspecto avaliado neste processo está relacionado à qualidade da sincronização das legendas com os eventos temporais da movimentação da boca nos quadros (sequência de imagens) dos vídeos, exigindo ajustes manuais e futuramente a necessidade de técnicas multimodais automatizadas para análise e sincronização destes eventos.

Em função destes resultados, faz-se necessário desenvolver métodos multimodais adicionais (áudio e imagem) para cortar de forma fidedigna e automática os instantes de início e fim dos sinais da "*palavra semente*" nas sentenças/frases encontradas nas ocorrências oriundas indicadas as legendas. Outra melhoria futura consiste em aumentar a qualidade dos resultados dos passos IV e V (Figura 2), a

qual está relacionada à sincronização de FPS (*frames per second*), uma vez que foi constatada uma variação entre 25 e 30 FPS nos vídeos utilizados - com tal sincronização, será possível garantir a representação correta de cada quadro durante o processo de comparação entre videos diferentes.

A partir da abordagem proposta neste trabalho é possível futuramente gerar um *dataset* pioneiro que represente medidas visuais da face ao longo do tempo com rotulagem auxiliada por legendas de vídeos públicos e que ao final, contemple uma capilaridade maior de variações de classes de sinais que representam uma mesma palavra seja no mesmo vídeo ou em vídeos de diferentes fontes.

5. CONCLUSÃO

Com a metodologia proposta foi possível prototipar a extração e normalização de medidas de boca que contribuem para habilitar a geração preliminar de assinaturas de movimentação labial e consequentemente, podem ser correlacionadas e rotuladas a partir da sincronização com as respectivas legendas dos vídeos.

Ainda há algumas limitações com relação a busca por similaridade, principalmente no que diz respeito às variações de formato e duração dos eventos da fala da palavra dentro de um mesmo vídeo e, no geral, tais variações se intensificam quando são tratados vídeos de origens diferentes. Tais aspectos implicam no uso adicional de outras técnicas, como machine learning e redes neurais, para realizar a classificação e busca das ocorrências cruzadas via sementes.

Vale destacar a necessidade de avaliar a qualidade da sincronização das legendas de forma multimodal, bem como verificar formas automatizadas de recortar as palavras nas sentenças/frases.

Como trabalhos futuros há a possibilidade de montagem de *datasets* no domínio das séries temporais extraídas a partir de vídeos reais rotulados por meio de legendas. Com a automatização do processo de extração visual das medidas e correlação com as legendas, será possível aumentar o volume de dados, habilitando novas análises no âmbito das redes que realizam classificação e, até mesmo, predição de eventos temporais. Com uma massa de dados maior será possível estudar outros efeitos como a subamostragem de movimentos em função das limitações de FPS de vídeos convencionais, em torno de 30 FPS.

Adicionalmente, com foco em aumentar a generalização das condições de identificação e diminuir as limitações causadas pela abordagem desta proposta poderão também ser explorados itens multimoldais complementares que combinam eventos temporais obtidos via as medidas em séries temporais deste trabalho com outros sinais temporais de outras fontes como por exemplo audio e bio-sensores.

Em resumo, é factível salientar que a partir deste trabalho foi possível extrair alguns resultados preliminares dos blocos 1 e 2 que precisarão de aperfeiçoamento. Foi verificado também que é possível relacionar as assinaturas dos sinais temporais medidos de forma visual com os efeitos de movimentação gerados pela fala. Como produto deste trabalho, os resultados preliminares serão publicados em um repositório público de modo a receber contribuições cruzadas de pesquisas similares da comunidade científica.

REFERÊNCIAS

Chasmai, M., Das, N., Bhardwaj, A., e Garg, R. (2022). A view independent classification framework for yoga postures. *SN Computer Science*, 3, 476. doi:10.1007/s42979-022-01376-7.

Chung, J.S. e Zisserman, A. (2018). Learning to lip read words by watching videos. *Computer Vision and Image Understanding*, 173, 76–85. doi:https://doi.org/10.1016/j.cviu.2018.02.001.

Dalvi, C., Rathod, M., Patil, S., Gite, S., e Kotecha, K. (2021). A survey of ai-based facial emotion recognition: Features, ml dl techniques, age-wise datasets and future directions. *IEEE Access*, 9, 165806–165840. doi:10.1109/ACCESS.2021.3131733.

Dupont, S. e Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3), 141–151. doi:10.1109/6046.865479.

Ekman, P., Friesen, W., e Hager, J. (2002). *Facial Action Coding System: Facial action coding system : the manual : on CD-ROM*. Facial Action Coding System. Research Nexus.

Fenghour, S., Chen, D., Guo, K., Li, B., e Xiao, P. (2021). Deep learning-based automated lip-reading: A survey. *IEEE Access*, 9, 121184–121205. doi:10.1109/ACCESS.2021.3107946.

Happy, S.L. e Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*, 6, 1–12. doi:10.1109/TAFFC.2014.2386334.

Kaynak, M., Zhi, Q., Cheok, A., Sengupta, K., Jian, Z., e Chung, K.C. (2004). Analysis of lip geometric features for audio-visual speech recognition. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 34(4), 564–570. doi:10.1109/TSMCA.2004.826274.

Kazemi, V. e Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874. doi:10.1109/CVPR.2014.241.

King, D.E. (2009). Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10, 1755–1758.

Ladani, D.J. e Desai, N.P. (2020). Stopword identification and removal techniques on tc and ir applications: A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 466–472. doi:10.1109/ICACCS48705.2020.9074166.

Law, S.M. (2019). Stumpy: A powerful and scalable python library for time series data mining. *Journal of Open Source Software*, 4(39), 1504. doi:10.21105/joss.01504.

Noceti, N., Sciutti, A., e Rea, F. (eds.) (2020). *Modelling Human Motion*. Springer Cham, 1 edition. doi:10.1007/978-3-030-46732-6. Published: 10 July 2020.

Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikäinen, M., e Liu, L. (2024). Deep learning for visual speech analysis: A survey.

Zhou, Z., Zhao, G., Hong, X., e Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9), 590–605. doi: https://doi.org/10.1016/j.imavis.2014.06.004.