Modelo de Classificação Automática de Frases Faladas com Abordagem em Redes Neurais Convolucionais

Cid Ivan C. Carvalho[1], Francisca Ticiany B. L. Oliveira[2], Vitória Maria A. Silva[2]

[1] Universidade Federal Rural do Semi-Árido UFERSA) - Mossoró, RN - Brasil
[2] Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró, RN - Brasil

cidivanc@gmail.com, ticianyoliveira@uern.br, vitoriamasmas@gmail.com

Abstract. *The article presents an automatic classification model of spoken sentences for Portuguese using convolutional neural networks (CNNs). The methodology involves the analysis of MFCC spectrograms as input to the CNN, treating the acoustic analysis. The model results are analyzed in terms of precision, recall, f-score, and accuracy for different categories. The study concludes that, although the model shows promising performance in some classifications, it still presents significant challenges in identifying canonical and anti-topic sentences, needing more audio data and future adjustments.*

Resumo. *O artigo apresenta um modelo de classificação automática de frases faladas para o português utilizando redes neurais convolucionais (CNNs). A metodologia envolve a análise de espectrogramas MFCCs como entrada para a CNN. Os resultados do modelo foram analisados em termos de precisão, recall, f-score e acurácia para diferentes categorias. O estudo conclui que, embora o modelo se mostre com desempenho promissor em algumas classificações, ele ainda apresenta desafios significativos na identificação de frases canônicas e anti topicalizadas, necessitando de mais dados de áudios e ajustes futuros.*

Palavras-chave: Modelo de classificação. Redes neurais. Frases faladas.

1. Introdução

A classificação automática de frases faladas pode contribuir na melhoria e no desenvolvimento de tecnologias de fala e em aplicações específicas em inteligência artificial como: assistentes virtuais - com síntese de fala mais próxima à prosódia humana e classificação de emoções - marcas sentimentais expressas na entoação da frase. Além disso, esses aplicativos geralmente se integram ao processamento de reconhecimento de fala, com a tarefa de marcar a pontuação no texto, e síntese de voz, na marcação de pausa, entoação, tons de fronteira. Para uma visão geral sobre o seu uso na tarefa de previsão de pontuação em contexto, consulte Casanova *et al*. (2024). A classificação de sentenças pode aprimorar significativamente a funcionalidade de sistemas de reconhecimento de fala. Vejamos alguns exemplos: o *Whisper* que realiza o reconhecimento de fala para algumas línguas, inclusive para língua portuguesa, o *Google Speech-to-Text API* e a biblioteca *Speech Recognition* para a linguagem Python.

Geralmente, os sistemas de reconhecimento de fala apresentam alta precisão na transcrição de segmentos fonéticos para os símbolos gráficos da língua portuguesa, por exemplo. No entanto, existem outras tarefas de fala que não são consideradas por esses sistemas quando realizam a transcrição da fala para a ortografia. Nesse contexto, entra a ideia de classificação de sentenças para vincular a tarefa de transcrição com as marcas de pontuação. Geralmente, os reconhecedores de voz não marcam, na escrita, elementos prosódicos que estão acima dos segmentos sonoros, como as pausas, os acentos de

pitch, a duração. Esses elementos têm relação direta com o uso do ponto final da frase, a vírgula, o ponto e vírgula e o ponto de interrogação, ou seja, marcas gráficas que não aparecem na transcrição automática de muitos sistemas de reconhecimento de voz. Essas marcas exigem que o sistema reconheça o padrão entoacional da fala após marcá-lo na escrita, como apontam Raso, Teixeira e Barbosa (2020).

Considerando esses aspectos, este trabalho tem o objetivo de apresentar um modelo de classificação automática de frases faladas com abordagem de redes neurais convolucionais. A classificação é feita com base na dimensão dos sintagmas que definem os sujeitos e os objetos, no tipo de verbo e na posição do objeto e do sujeito nas frases. Destacamos que este trabalho apresenta um protótipo de um modelo probabilístico para classificação automática de frases. Para isso, separamos gravações de frases declarativas produzidas na variação da língua portuguesa falada no estado do Rio Grande do Norte. Esses dados serviram de base para a criação do modelo desenvolvido em redes neurais convolucionais, como pode ser visto nos tópicos a seguir.

2. Fundamentação teórica

A língua portuguesa é classificada como sendo uma língua de sujeito e predicado, ou seja, a estrutura oracional segue uma sequência em que o sujeito surge primeiro, seguido do verbo e por último o objeto (SVO), que chamaremos de *ordem canônica* da frase. No entanto, há autores que discutem a rigidez com que a classificam, uma vez que essa língua pode apresentar outras ordens na estrutura sintática sem a perda do sentido. De fato, é efetiva e predominante a ordem direta (SVO), todavia, as relações sintáticas com estruturas topicalizadas e anti topicalizadas são amplamente usadas, em especial, na modalidade falada. Como bem aponta Pontes (1987), o português sempre foi considerado uma língua de predominância sujeito-predicado. No entanto, o uso cotidiano da língua pelos falantes evidencia a existência de outras estruturas, como as de tópico-comentário. Quando o objeto é deslocado para a posição esquerda da sentença, o sintagma nominal "ocupa uma posição não argumental, externa à sentença e simétrica àquela dos tópicos". [Berlink, Duarte e Oliveira, 2017. p. 129]. Desse modo, os constituintes podem se movimentar na sentença, criando outras estruturas, mas com a mesma correspondência semântica, no entanto, nem sempre com a mesma ênfase e/ou foco prosódico.

A diferença entoacional também se caracteriza na mudança de posição dos sintagmas e indicam que o sujeito ou o objeto foi deslocado. Nesse sentido, os conceitos que norteiam este trabalho são: a entoação, a frequência fundamental (doravante F0), a duração e a intensidade. O conceito de entoação é fundamental para compreendermos a diferença na organização e na combinação dos elementos sintáticos, pois ele revela a marcação de graves e de agudos ao longo da cadeia da fala, como dizem Barbosa (2019, 2022) e Lucente (2022). A F0 corresponde ao número de vezes que as pregas vocais completam ciclos de vibrações em intervalos regulares de um segundo [Barbosa 2019], sendo considerada o principal correlato acústico para a determinação do padrão entoacional. Outro correlato acústico importante é a duração, que indica a extensão de tempo envolvida na articulação de um som ou sílaba. É medida em milissegundos, quando relacionada a unidades menores que a palavra e em segundos, quando relacionada à palavra e à sentença. Por último, a intensidade é um correlato que expressa "o quão forte um som é" [Barbosa 2019, p. 26]. Ela se relaciona ao tamanho da resistência que a glote oferece à passagem do ar, também com a quantidade e a

velocidade com que o ar atravessa, somadas à pressão sofrida pelo tamanho das pregas vocais.

Além desses conceitos envolvidos na implementação do modelo, para a classificação automática dos tipos de frases, utilizamos os métodos de aprendizado de máquina da rede neural convolucional. Segundo Catarino (2025), esse tipo de rede é aplicado na classificação de imagens, mas também na classificação de áudio. Esse método utiliza quatro etapas para identificar os padrões de áudio: a primeira é a camada de convolução, que cria os detectores de aspectos das imagens; a segunda etapa é a aplicação da camada *Max Pooling*, que reduz as dimensões das imagens; a terceira é a camada de achatamento, que transforma a matriz em um vetor; e a última etapa é a camada de conexão completa, veja [Srivastava 2014]. Para identificação dos áudios, utilizamos como entrada do sistema as imagens do espectrograma, que correspondem a "uma representação visual do espectro de frequência de um sinal de áudio ao longo do tempo" [Casanova 2024]. Para o processamento do áudio nos métodos de redes neurais artificiais, utilizamos os espectrogramas de Coeficientes Cepstrais de Frequência Mel (MFCC). A escala Mel é um tipo de escala com transformação não linear que converte a banda de frequência do áudio em uma banda de áudio que soa idêntica ao ouvido humano. Em outras palavras, a transformação do som de igual distância na escala mel é obtida como visto de igual distância para o humano, para mais informações veja [Wibawa e Darmawan 2021].

Além desses conceitos, apontamos alguns trabalhos que descrevem o processo de rotulagem e detecção de elementos prosódicos. A pesquisa de Wightman e Ostendorf (1994) descreve um algoritmo que utiliza árvores de decisão e um modelo de sequência de Markov para definir os padrões prosódicos, como fraseamento e proeminência, após o reconhecimento de palavras. Wagner (2008) propõe uma estrutura para a rotulagem automática da prosódia focada na detecção de sílabas acentuadas, limites de frases e no reconhecimento de acentos de altura e tons de contorno, utilizando modelos de classificação baseados em vetores de características acústicas. A pesquisa de Raso, Teixeira e Barbosa (2020) investiga os parâmetros fonético-acústicos na percepção de fronteiras prosódicas na fala espontânea do português brasileiro, desenvolvendo modelos para a detecção automática dessas fronteiras, com especial atenção à influência das pausas físicas. Esses trabalhos mostram que a classificação automática de frases faladas se constitui como um estudo fundamental no reconhecimento de voz.

3. Metodologia

Como mencionamos na seção anterior, o método de análise acústica foi tratado como um problema de classificação de imagens, onde o espectrograma representa uma imagem visual do som. Essa abordagem dispensa a extração manual de parâmetros acústicos, como frequência fundamental (F0), duração e intensidade, permitindo que o modelo aprenda as características dos sinais da fala que são mais salientes para diferenciar as estruturas linguísticas. No entanto, os aspectos dos áudios ficam destacados em outro nível de informação que não correspondem a dados acústicos, claro que sem a perda da qualidade das informações. Com essa perspectiva, construímos um conjunto de dados composto por 612 frases declarativas balanceadas, com apenas um canal (mono) no formato de arquivo com extensão *.wav* e com duração média de 1,8 segundos, apresentando uma taxa de amostragem de 44.100 Hz e ocupando um espaço

em disco de 248,8 megabytes. Todas as sentenças foram gravadas em laboratório com a aprovação do conselho de ética (CAAE 79118423.5.0000.9547). Para a gravação, foram convidados 20 participantes brasileiros, falantes da língua portuguesa, com faixa-etária de idade compreendida entre 18 a 40 anos, sendo dez do sexo masculino e dez do feminino.

Para a coleta dos dados dos áudios, os participantes foram posicionados em frente a tela do *notebook* com dimensão de tela de 17 polegadas, no qual se encontrava instalado o programa *Psychopy*. Eles utilizaram um microfone portátil de lapela conectado a um *smartphone* da *Apple* no qual tinha sido instalado um aplicativo de gravação chamado Gravador Fácil. Após a produção de uma frase, o participante avançava para a próxima tela pressionando a tecla ”espaço”. As gravações ocorreram em ambiente controlado, livre de ruídos externos, para minimizar interferências e preservar a qualidade do sinal. Cada participante produziu as sentenças experimentais três vezes. Para este protótipo inicial, foram selecionados 612 áudios. Todos os arquivos, bem como os códigos utilizados no *Google Colab*, estão disponíveis no repositório *GitHub* [link]. Na etapa seguinte, elaboramos uma função de pré-processamento para extrair a taxa de amostragem e o número de amostras, gerando espectrogramas MFCCs por meio da biblioteca Librosa. Essas imagens foram passadas como parâmetros para a biblioteca *Numpy* a qual transforma os dados em *array*. Depois dessa etapa, separamos os dados de treino e de teste, sendo que 80% (326) dos áudios para treino e 20% (82) áudios para o teste. O modelo foi treinado por 50 épocas, com validação ao final de cada uma. No tópico a seguir, apresentaremos análise e resultados que obtivemos com a criação do modelo.

4. Descrição e análise dos resultados

Para a descrição e análise dos resultados, organizamos as informações em duas etapas. Na primeira, examinamos as métricas de precisão, *recall*, *f-score* e acurácia do modelo na classificação de sentenças faladas, considerando a dimensão dos sujeitos e objetos, o tipo de verbo e a posição do sujeito e do objeto. A Tabela 01 apresenta os valores obtidos para cada uma das classes utilizadas na classificação das sentenças.

Tabela 01. Análise do modelo para a classificação das sentenças faladas

Classes	Precision	Recall	F-score	Accuracy
Sujeito e objeto complexos	0.90	0.77	0.83	0.82
Sujeito e objeto simples	0.75	0.89	0.81	
Verbo transitivo direto	0.85	0.85	0.85	0.89
Verbo transitivo indireto	0.91	0.91	0.91	
Anti topicalizada	0.59	0.55	0.57	0.60
Canônica	0.58	0.74	0.65	
Topicalizada	0.68	0.49	0.57	

Fonte: Dados da pesquisa

A (Tabela 01) mostra que, quando o modelo classifica uma frase como tendo sujeito e objeto complexos, ele apresenta uma precisão de 0.90; o *recall* de 0.77 mostra

que ele identifica 77% das frases com sujeito e objeto complexos; o *f-score* de 0.83 indica que o modelo tem algumas dificuldades em classificar categoria. Já para o sujeito e objeto simples, o modelo tem uma precisão de 0.75; o *recall* do modelo é de 0.89 e o *f-score* 0.81, o que indica que há um equilíbrio entre precisão e *recall*, apesar de uma precisão mais baixa. Na classificação do verbo transitivo direto, tem uma precisão de 0.85. No entanto, ressalta-se que a quantidade de frases utilizada para do modelo ainda é restrita para tirar conclusões precisas sobre esse fato; o *recall* para essa classe apresentou um valor de 0.85; o *f-score* foi de 0.85, talvez impulsionado pela precisão. Para a classe dos verbos transitivos indiretos, a precisão na classificação foi de 0.91; o *recall* tem um valor de 0.91, mostrando que o modelo identifica as frases que realmente contêm um verbo transitivo indireto. No entanto, da mesma forma que ocorreu com a precisão do verbo direto, deve-se ressaltar que a quantidade de frases utilizada ainda é limitada. O *f-score* foi de 0.91, este é o maior *f-score* da tabela, indicando um desempenho excepcional para esta categoria.

A classificação quanto à posição do objeto e do sujeito nas frases é a que apresenta maior dificuldade. Para a classe das frases anti topicalizadas, a precisão foi de 0.59, mostrando que o modelo tem muita dificuldade em acertar quando a frase é anti topicalizada, pois apresenta muitos falsos positivos. O *recall* de 0.55 mostra que o modelo identifica apenas cerca de metade dos casos reais dessa categoria. Isso reflete diretamente no *f-score* com 0.57, o desempenho geral para essa categoria é fraco. Para a classe das frases canônicas, a precisão foi de 0.58, o modelo ainda erra muito quando prevê esta categoria; o *recall* foi de 0.74, mostrando que o modelo só consegue identificar 74% dos casos reais de frases canônicas; o *f-score* de 0.65 foi a categoria com pior desempenho. Para a classe das frases topicalizadas, o modelo apresenta uma precisão mediana de 0.68; o *recall* acompanha o valor da precisão com o valor de 0.49; o *f-score* do modelo foi de 0.57 com um desempenho abaixo da média, mas com valor maior que para a classificação das frases anti topicalizadas e canônicas. A acurácia do modelo para a classificar a dimensão dos sujeitos e dos objetos foi de 0.82, mostrando que apresenta acerto em apenas 82% da entrada e comete 18% de erro. A acurácia de 0.89 para as categorias de verbos é boa. Já na classificação da posição de sujeito e objeto nas sentenças é muito baixa 0.60. Espera-se o aumento da acurácia do modelo com o acréscimo dos dados.

Feita essa análise, passamos à segunda etapa, que consiste na avaliação com base na quantidade de verdadeiros positivos e falsos negativos. Para a classe sujeito e objeto complexos, o modelo apresentou taxa de acerto de 29 casos, mas incorreu em aproximadamente 17 erros, nos quais confundiu tais estruturas com a classe sujeito e objeto simples. Em outras palavras, nessas 17 ocorrências, o modelo simplificou a classificação, deixando de reconhecer a complexidade presente. Por outro lado, para a classe sujeito e objeto simples, o desempenho foi superior: 29 acertos e apenas 7 erros, nos quais frases simples foram classificadas como complexas. Esses resultados indicam que o modelo é mais confiável na identificação de estruturas simples do que na detecção de estruturas complexas.

Quando da posição do sujeito e do objeto na frase, o modelo identifica a classe das sentenças canônicas com maior sucesso. Ele acerta 34 vezes e comete apenas 12 erros, sendo que 9 com a classe anti topicalizada e 3 com a topicalizada, mostrando certa capacidade de reconhecer esta categoria. Por outro lado, a classe anti topicalizada

foi a classe que apresentou maior dificuldade em classificar. Ele acerta apenas 23 vezes e comete 19 erros. Isso mostra que há confusão com as outras duas classes, especialmente, com a canônica, que contabilizou 14 erros, e topicalizada, com 5 erros. Dessa forma, entende-se que o modelo identifica corretamente as frases anti topicalizadas. Já a classe anti topicalizada apresenta um desempenho inferior em relação às duas classes anteriores. O número de acertos é de 17 frases menor que a soma dos erros das classes anteriores: 18, mas ainda há uma confusão considerável e distribuída igualmente entre as outras duas classes. O modelo ainda precisa de ajuste nos parâmetros para identificar a classe topicalizada como não sendo canônica, indicando que ele não aprendeu as características distintivas dessa categoria. Por isso, esta será a área que exigirá maior atenção para as próximas implementações do modelo. Outro ponto que merece atenção é a classe anti topicalizada, a qual é frequentemente confundida com as outras, embora com menos intensidade que a canônica. Quando se analisa a classe do verbo transitivo direto, o modelo acertou 39 vezes e errou 7 vezes, classificando como verbo transitivo indireto. Isso indica que o sistema apresenta pouca dificuldade em identificar corretamente os verbos transitivos diretos, ou confunde em alguns deles com os indiretos. Em relação ao verbo transitivo indireto, o modelo foi mais eficaz do que para a classe anterior, acertando 70 vezes. O modelo apresentou um desempenho bom na classificação de verbos transitivos indiretos. No entanto, a classificação para verbos transitivos diretos indica que o sistema tem dificuldade em reconhecer todos os verbos.

5. Conclusão

A implementação do modelo acima demonstrou que há uma complexidade inerente ao processamento computacional da linguagem falada, especialmente, no que se refere à prosódia da língua. Embora a abordagem seja tecnicamente viável, os resultados obtidos revelaram algumas falhas significativas, com o modelo apresentando uma alta taxa de erro e dificuldade na classificação das categorias propostas no posicionamento do objeto ou do sujeito. No entanto, apesar do desempenho para essas classes, este trabalho não deve ser visto como um insucesso, mas sim como um diagnóstico inicial e um ponto de partida para integração aos reconhecimento de voz.

Como visto, o modelo tende a cometer erros bem específicos ao classificar frases relacionadas a posição do sujeito e do objeto, todavia, o erro no sentido inverso apresenta-se menos frequente. Nesse aspecto, ele apresenta uma regular capacidade de identificar corretamente as duas classes quando a previsão está correta. Um ponto que destacamos para os próximos treinos é reduzir a quantidade de falsos negativos para essa classe do modelo. Verificamos que há um desempenho bastante desigual entre as classes do modelo. Enquanto, o modelo é satisfatório com a classe topicalizada, ele falha drasticamente com a classe não topicalizada, tornando-o pouco confiável para essa tarefa específica. Por isso, os passos seguintes envolvem acréscimos no número de amostras de treinamento para esta classe específica, ajustes de parâmetros das redes neurais, revisão das características usadas pelo modelo, aplicação de técnicas de balanceamento das classes, uma vez que os dados originais tem menos exemplos de frases anti topicalizadas do que as outras.

Referências

Barbosa, P. A. (2019). *Prosódia*, São Paulo, Parábola.

Barbosa, P. A. (2022). *Manual de Prosódia Experimental*, 1 ed, Editora da Abralin.

Berlinck, R. A.; Duarte, M. E. L. and Oliveira, M. (2009). Predicação. In Castilho, A. T., Kato, M. A., Nascimento, M., *Gramática do português culto falado no Brasil*, Campinas, Editora da Unicamp.

Casanova et al. (2024). Recursos para o Processamento de Fala. In Caseli, H. M. and Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 2 ed, BPLN. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao.

Catarino, M. H (2025). Redes Neurais. Rio de Janeiro, Editora Freitas Bastos.

Lira, Z. (2009). A entoação modal em cinco falares do Nordeste brasileiro, Tese (Doutorado) - UFPB, João Pessoa.

Lucente, L. (2022). Notação Entoacional. In Oliveira Júnior, M. *Prosódia, Prosódias*, Editora Contexto, pages 45-66.

Pontes, E. (1987). O Tópico no Português do Brasil. Campinas, Editora Pontes.

Raso, T.; Teixeira, B.; Barbosa, P. (2020). Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. Journal of Speech Sciences, Campinas, SP, v. 9, n. 00, pages 105–128. DOI: 10.20396/joss.v9i00.14957. Disponível em: https://econtents.bc.unicamp.br/inpec/index.php/joss/article/view/14957.

Srivastava, N. et al. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, pages 1929-1958. https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

Wibawa, I.D.G.Y.A; Darmawan, I.D.M.B.A (2021). Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini, Journal of Physics: Conference Series, DOI: 10.1088/1742-6596/1722/1/012014.

Wightman, C.W.; Ostendorf, M. (1994). Automatic labeling of prosodic patterns, EEE Transactions on Speech and Audio Processing, 2(4), pages 469–481.

Wagner, A. (2008). Automatic labeling of prosody, ITRW on Experimental Linguisriucs, ExLing 2008, 25-27 August 2008, Athens, Greece. Disponível em: https://www.isca-archive.org/exling_2008/wagner08_exling.pdf.