Classificação automática de sentenças faladas com arquitetura híbrida: análise de estruturas topicalizadas, anti-topicalizadas e canônicas

Gerson Vitor P. Fernandes[1], Vitoria Maria A. Silva[2], Cid Ivan C. Carvalho[3]

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

prof.gersonvitor@gmail.com , cidivanc@ufersa.edu.br, vitoriamasmas@gmail.com

Abstract. *This article investigates the performance of an automatic classification system for spoken Brazilian Portuguese sentences considering canonical, topicalized, and antitopicalized structures using a hybrid architecture based on Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). The research integrates syntactic, prosodic, and computational contributions to contextualize the corpus and define the scope of sentence processing; However, the study focuses on the experimental evaluation of a hybrid neural network architecture applied to the automatic classification of these sentences. The analyzed corpus consists of 612 declarative sentences recorded in a controlled environment, produced by 20 Brazilian Portuguese speakers. The methodology is quantitative and experimental, employing supervised learning and evaluation based on metrics such as precision, recall, F-score, and accuracy, using a confusion matrix. Tests were conducted with various hyperparameter configurations, varying convolutional filters, LSTM units, dropout rates, and the number of training epochs. The results demonstrated that more robust models exhibited a greater capacity to memorize training data but also a higher tendency toward overfitting. Conversely, leaner architectures combined with higher regularization rates achieved better generalization capabilities and greater stability during validation. The best performance was observed in the scenario with 30% dropout, which showed a significant reduction in the generalization gap between training and testing. It is concluded that the combination of RNNs and CNNs shows promise for the automatic classification of spoken sentences, although limitations regarding model generalization to unseen data still exists.*

Keywords: *Classification model. Neural networks. Convolutional neural networks. Spoken sentences.*

Resumo. *O presente artigo investiga o desempenho de um sistema de classificação automática de sentenças faladas do português brasileiro, considerando estruturas canônicas, topicalizadas e antitopicalizadas, por meio de uma arquitetura híbrida baseada em Redes Neurais Recorrentes (RNNs) e Redes Neurais Convolucionais (CNNs). A pesquisa articula contribuições sintáticas, prosódicas e computacionais com a finalidade de contextualizar o corpus e delimitar o tratamento das sentenças faladas, porém o estudo concentra-se na avaliação experimental de uma arquitetura híbrida de redes neurais aplicada à classificação automática dessas sentenças. O corpus analisado é composto por 612 sentenças declarativas gravadas em ambiente controlado, produzidas por 20 falantes do português brasileiro. A metodologia caracteriza-se como quantitativa e experimental, utilizando aprendizagem supervisionada e avaliação baseada em métricas como precisão, recall, f-score e acurácia, por meio da matriz de confusão. Foram realizados testes com diferentes configurações de hiperparâmetros, variando filtros convolucionais, unidades LSTM, taxas de Dropout e número de épocas de treinamento. Os resultados demonstraram que modelos mais robustos apresentaram maior capacidade de memorização dos dados de treino, mas também maior tendência ao overfitting. Em contrapartida, arquiteturas mais enxutas associadas a maiores taxas de regularização obtiveram melhor capacidade de generalização e maior estabilidade durante a validação. O melhor desempenho foi*

observado no cenário com 30% de Dropout, em que houve redução significativa do hiato de generalização entre treino e teste. Conclui-se que a combinação entre RNNs e CNNs mostra-se promissora para a classificação automática de sentenças faladas, embora ainda existam limitações relacionadas à generalização do modelo em dados não vistos.

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

1. Introdução

Nas últimas décadas, os avanços no campo da Linguística Computacional e da Inteligência Artificial têm possibilitado o desenvolvimento de sistemas capazes de processar, reconhecer e classificar dados linguísticos de forma automatizada. Entre essas possibilidades, destaca-se o processamento automático da fala, área que reúne contribuições da linguística e da computação para a investigação de padrões linguísticos por meio de modelos computacionais treinados para reconhecimento e classificação de estruturas da língua.

Em estudos que envolvem dados de fala, aspectos fonético-fonológicos e prosódicos também podem ser incorporados à análise, especialmente quando se investigam relações entre entoação, organização sintática e processamento acústico. No âmbito da fala, a prosódia exerce papel fundamental na organização e interpretação dos enunciados, uma vez que elementos como entoação, duração, intensidade e ritmo contribuem para a construção de sentidos e para a distinção entre diferentes configurações sintáticas.

No português brasileiro, além das estruturas consideradas canônicas pela tradição gramatical, observa-se ampla produtividade de construções topicalizadas e antitopicalizadas, especialmente na modalidade oral, evidenciando a flexibilidade organizacional da língua e a relação entre sintaxe e prosódia. Os estudos sobre topicalização no português brasileiro ganharam destaque a partir das discussões de Pontes (1987), que apontam a coexistência de estruturas orientadas tanto para o sujeito quanto para o tópico. Posteriormente, pesquisas como as de Berlinck, Duarte e Oliveira (2009) ampliaram essa discussão ao sistematizar diferentes tipos de construções de tópico marcado. Paralelamente, investigações recentes têm buscado compreender de que maneira tais estruturas podem ser identificadas automaticamente por sistemas computacionais, sobretudo a partir de dados de fala.

Nesse contexto, o uso de redes neurais artificiais tem se mostrado promissor em tarefas relacionadas ao reconhecimento de padrões linguísticos e acústicos. Modelos baseados em aprendizado de máquina vêm sendo empregados em diferentes aplicações envolvendo a fala, como transcrição automática, detecção de distúrbios vocais e classificação de sentenças faladas. Entre essas arquiteturas, as Redes Neurais Recorrentes (RNNs) destacam-se pela capacidade de lidar com informações sequenciais e temporais, característica particularmente relevante para o processamento de sinais de fala.

Diante desse cenário, o presente estudo busca responder à seguinte questão: considerando as estruturas sintáticas canônicas, topicalizadas e anti-topicalizadas, quais apresentam melhor e pior desempenho em um sistema de classificação automática de sentenças faladas baseado em redes neurais recorrentes? Diante dessa questão, tem-se o seguinte objetivo geral: analisar o desempenho da classificação automática de frases faladas por meio de uma abordagem híbrida de redes neurais, considerando as categorias topicalizadas, anti-topicalizadas e canônicas.

Para isso, utiliza-se um corpus composto por sentenças declarativas do português brasileiro previamente gravadas em ambiente controlado e organizadas para

treinamento e validação do modelo computacional. A pesquisa articula três interfaces: sintaxe, prosódia e computação, buscando contribuir tanto para os estudos linguísticos acerca da organização sentencial do português brasileiro quanto para o desenvolvimento de modelos automáticos de classificação da fala.

Após esta introdução, apresentam-se os pressupostos teóricos que fundamentam a pesquisa, contemplando discussões acerca das construções topicalizadas, da prosódia e das redes neurais recorrentes e redes neurais convolucionais. Em seguida, descrevem-se os procedimentos metodológicos adotados para a constituição do corpus, treinamento do modelo e análise dos resultados.

2. Fundamentação teórica

2.1 Sentenças topicalizadas, anti-topicalizadas e canônicas

Em uma perspectiva normativa do português brasileiro, é consensual a adoção da ordem *sujeito + verbo + objetos + complementos* como o padrão organizacional mais comum dos elementos sentenciais. No entanto, na modalidade oral, outros padrões organizacionais são igualmente produtivos, como é o caso das sentenças de tópico. De acordo com Kenedy (2014), essas estruturas ocorrem quando um determinado constituinte (o tópico) é posicionado à periferia esquerda de uma sentença e, após essa colocação, há um comentário sobre esse elemento, como é o caso do seguinte exemplo: "*o vestido vermelho, a moça de preto comprou",* em que o objeto direto da sentença é posto na posição de tópico e o restante da sentença é organizado posteriormente. Desse modo, compreende-se que os constituintes podem se movimentar na sentença, criando outras estruturas, com a mesma correspondência semântica, no entanto, nem sempre com a mesma ênfase e/ou foco prosódico (Carvalho, Oliveira e Silva, 2025).

A discussão sobre a produtividade dessas construções teve início a partir de Li e Thompson (1976), que classificaram as línguas naturais em quatro tipologias básicas, a saber: i) línguas orientadas para a sentença - com proeminência de sujeito; ii) línguas orientadas para o discurso - com proeminência de tópico; iii) línguas com proeminência de tópico e sujeito; e iv) línguas que não têm proeminência de tópico nem de sujeito (sincréticas). No Brasil, as estruturas topicalizadas ganharam notoriedade com os estudos apresentados por Pontes (1987), que classificaram o português como uma língua de tópicos e sujeitos, haja vista a produtividade das sentenças topicalizadas.

Segundo Berlinck, Duarte e Oliveira (2009), as construções de tópico marcado não formam um grupo homogêneo. Com isso, é possível classificar o tópico marcado em, pelo menos, cinco tipos: 1) tópico pendente; 2) deslocamento à esquerda; 3) topicalização; 4) tópico-sujeito e 5) antitópico. No primeiro tipo, há conectividade semântica entre os constituintes e a sentença-comentário, mas não há conectividade sintática, como em "*drama, já basta a vida"*. No segundo tipo de tópico, um constituinte na posição de tópico tem um correferente expresso na sentença comentário, como em "cada elemento, *ele* possui o seu conjunto". O terceiro tipo envolve o movimento de um constituinte de sua posição de origem para a posição de tópico, deixando a posição original "vazia", como em "*carne, aqui em casa nós fazemos de várias formas",* em que o objeto direto é deslocado para a posição de tópico. O quarto tipo é aquele em que o tópico se encontra numa sentença com um sujeito nulo não-argumental, como em "meu celular acabou a bateria", em que parte de um constituinte é deslocado para a posição topicalizada. O último tipo de tópico apresentado é o antitópico que aparece à direita da sentença e ocupa uma posição

não-argumental e externa, como em “leva azeite de dendê, o acarajé”.

Neste artigo, o modelo de classificação considerou três possibilidades de organização, a saber: a organização considerada canônica pela tradição gramatical e pelos estudos linguísticos (sujeito + verbo + objeto); a segunda é a topicalização e a antitopicalização, na perspectiva de Berlinck, Duarte e Oliveira (2009).

Nesta seção, foi apresentado o conceito de tópico e suas diferentes classificações. Na seção seguinte, serão apresentados os conceitos relacionados à prosódia.

2.2 Prosódia

O termo “prosódia” deriva do grego προσῳδία, no qual o prefixo προσ significa "sobre" e a raiz ᾠδή, "canção". Originalmente, conforme explicam Rebollo-Couto e Seara (2019), o termo referia-se à "canção sobre os sons segmentais", sendo utilizado para descrever os acentos das palavras. Com o tempo, seu significado foi sendo expandido e passou a abranger uma gama mais ampla de elementos ligados à fala humana.

Assim, a prosódia passa a compreender um conjunto de fenômenos que envolvem parâmetros como altura, intensidade, duração, pausas e velocidade da fala, o estudo da entoação, acento e ritmo nas línguas naturais (Scarpa, 1999). De maneira geral, entende-se que a prosódia não está relacionada diretamente ao conteúdo da fala, mas ao modo “como” a fala é produzida, refletindo a maneira de transmitir significados por meio da variação sonora (Barbosa, 2019).

No estudo de Carvalho, Oliveira e Silva (2025), foi utilizada uma perspectiva prosódico-sintática, em que se observou a estrutura sintática e a sua realização na fala. Nesse estudo, elementos como a frequência fundamental (F0) e o contorno entoacional foram analisados, a fim de identificar diferenças na região nuclear do enunciado; o estudo apontou diferenças no contorno entre falantes masculinos e femininos.

O presente estudo articula-se entre três interfaces: 1. sintaxe, 2. prosódia e 3. computacional. A sintática relaciona-se pelas construções trabalhadas na seção 2.1: canônicas, topicalizadas e anti topicalizadas. O nível prosódico justifica-se pelo fato de trabalharmos com sentenças faladas. Embora a proposta deste estudo não seja realizar uma análise acústica dessas sentenças, elas foram gravadas em local controlado, devidamente tratadas e segmentadas, a fim de garantir a melhor qualidade para a análise computacional. Por fim, a interface computacional justifica-se por um sistema de classificação, adotando um modelo híbrido, conforme discutido na seção 2.3.

2.3 Classificação automática e redes neurais recorrentes

No campo da linguística computacional, o processamento automático da fala tem sido investigado em diálogo com a engenharia computacional, principalmente nos estudos envolvendo *o aprendizado de máquina* e *processamento de sinais*. Com base no estudo de Almeida *et al.* (2024), nota-se que os modelos treinados podem ser usados como uma ferramenta de transcrição da fala de indivíduos com gagueira. De forma semelhante, Azevedo (2025) em sua análise na identificação de distúrbios da voz, entre eles a *disfonia, laringite, pólipo e paralisia*, o autor apresentou dados positivos, no teste de validação, o modelo atingiu uma taxa de 90%, mesmo com algumas limitações.

Entretanto, através de estudos computacionais é possível realizar uma abordagem sintática, linguística e computacional, como no caso de Carvalho, Oliveira e Silva (2025), quel, baseados no modelo de redes neurais convolucionais, estudaram a classificação automática de sentenças faladas, através de espectrogramas. Mesmo alcançando uma acurácia abaixo dos 90%, o modelo apresentou uma excelente precisão em alguns tipos de sentenças.

O aprendizado de máquina parte do princípio que as máquinas conseguem aprender como uma "mente" humana, mesmo produzida artificialmente. É importante ressaltar que ainda existem muitas limitações, porém os modelos computacionais conseguem atingir excelentes resultados. Dijkstra (2021) cita três tipos de aprendizagem em seu estudo: o *supervisionado*, *não supervisionado e o reforço*. Cada treinamento tem uma abordagem específica, portanto, é necessário entender para qual finalidade o modelo será usado, a fim de encontrar o treino ideal.

No primeiro caso mencionado, o sistema aprende através de um banco de dados previamente "alimentado". A máquina tenta encontrar padrões, e, quanto mais dados tiver para trabalhar, mais fácil o reconhecimento. O segundo caso tenta encontrar e criar grupos de dados; mesmo não existindo um direcionamento inicial, ele analisa as características de cada uma, identifica padrões e agrupa no local que mais se aproxima com o objeto. O terceiro caso é baseado em um sistema de erros e acertos, recompensas e punições em um determinado ambiente; através de falso positivo e falso negativo, trata-se de um modelo que se ajusta de acordo com um direcionamento.

Assim como os tipos de aprendizado, percebe-se a necessidade de distinguir a diferença entre redes neurais recorrentes e convolucionais, pois cada rede tem uma abordagem diferente. Faceli (2011) explica que as redes neurais aproximam-se de uma tentativa artificial de simular como o sistema neural funciona biologicamente, sendo uma tentativa de automatizar tarefas cotidianas por meio de máquinas treinadas, que incluem neurônios artificiais. O conjunto deles forma as redes neurais artificiais (RNAs).

As redes neurais funcionam com arquiteturas diferentes, cada uma delas podendo ser mais adequada para tarefas específicas, como no caso das RNNs que trabalham com sequência e temporalidade, diferente das CNNs que conseguem encontrar padrões em múltiplas categorias como imagens, representações acústicas, espectrogramas, matrizes e outros.

No presente estudo, utilizamos uma abordagem híbrida usando as Redes Neurais Recorrentes (RNNs) e Redes Neurais Convolucionais para o sistema de classificação de sentenças faladas. Em primeiro caso, o estudo de Moura (2018), afirma que as redes neurais recorrentes são formadas por uma arquitetura de redes em camadas, que funciona da seguinte forma: inicialmente o sistema captura as relações temporais por meio do compartilhamento de parâmetros e a arquitetura considera as informações anteriores, em diferentes momentos do processamento sequencial, auxiliando na predição.

Por outro lado, as convolucionais distinguem-se por conseguir identificar padrões em dados estruturados em matrizes, imagens, espectrogramas ou até mesmo representações acústicas da fala. Contudo, sentenças faladas não são analisadas de forma bruta: é necessário transformá-las em alguma das variáveis mencionadas anteriormente, por meio de representações visuais ou numéricas. Conforme Dijkstra (2021), esse tipo de arquitetura é formada por diversas camadas, entre elas as de convolução e pooling.

É importante ressaltar que a forma de aprendizado que mais se aproxima com a nossa pesquisa é a aprendizagem supervisionada, pois o nosso objetivo é a identificação de três categorias sintáticas com características específicas, por meio do treinamento do sistema, espera-se resultados satisfatórios. Por fim, utilizamos uma abordagem híbrida, com um modelo de classificação baseado no sistema de redes neurais recorrentes (RNNs) e de redes neurais convolucionais (CNNs).

3. Metodologia

Como mencionado anteriormente, o estudo não concentra-se em pistas prosódicas, como duração, pausa, sílaba tônica e ou elementos acústicos, porém, por se tratar de sentenças faladas, adotamos todos os princípios de coleta e qualidade usados em análises prosódicas, incluindo o rigor exigido para a escolha dos falantes, instrumentos, comitê de ética (CAAE 79118423.5.0000.9547) e ambiente controlado.

O corpus utilizado é o mesmo conjunto adotado por Carvalho, Oliveira e Silva (2025), com um total de 612 sentenças declarativas, arquivos gerados no formato .wav. No total, foram 20 participantes, entre eles 10 do gênero masculino e 10 do feminino, ambos com faixa etária entre 18 a 40 anos. Foram selecionados entrevistados que não tivessem dificuldades de fala, causadas por questões cognitivas ou por uso de aparelhos ortodônticos, critérios adotados por questões metodológicas.

Como exemplo das sentenças inseridas no modelo, tem-se, na estrutura canônica, "O diretor concordou com a reunião", topicalizada "Com a reunião, o diretor concordou" e antitopicalizada "Concordou com a reunião, o diretor".

A pesquisa enquadra-se como um estudo quantitativo experimental, por utilizarmos uma boa quantidade de sentenças, como também um sistema pelo Google Colab com arquitetura híbrida para discutir a quantidade de erros e acertos do sistema.

A métrica para validação do sistema foi a matriz de confusão, uma tabela que consegue avaliar o desempenho de um sistema, por meio das variáveis: *precisão, recall, f-score e acurácia.* Com esses parâmetros é possível definir a qualidade da arquitetura, entretanto, para encontrar os valores dos elementos mencionados anteriormente, trabalha-se por meio de quatro possibilidades. Segundo Monard e Baranauskas (2003, p. 48) as possibilidades são: "*Verdadeiros positivos (Tp), Falsos positivos (Fp), Falsos negativos (Fn) e Verdadeiros negativos (Tn)".* O primeiro caso ocorre quando o sistema entrega o resultado corretamente; o segundo entrega um resultado, porém falso, não condizente com a verdade; o terceiro, quando é esperado um resultado verdadeiro e ele entrega um negativo; o quarto ocorre quando é esperado uma informação negativa e ele realmente entrega.

4. Descrição e análise dos resultados

Neste estudo, foi adotado o termo época referente a cada ciclo completo de treinamento do modelo usando os dados disponíveis. Os hiperparâmetros são configurações feitas antes do treinamento, como filtros, unidades LSTM, taxa de Dropout e número de épocas. Inicialmente, a avaliação do modelo híbrido baseou-se na variação controlada dos hiperparâmetros de convolução, memória e regularização. Foram definidos dois cenários de testes principais (A e B). Para cada um deles, aplicou-se uma variação na taxa de Dropout nos valores de 20%, 25% e 30%. A estrutura detalhada de cada cenário encontra-se resumida na **tabela 1**.

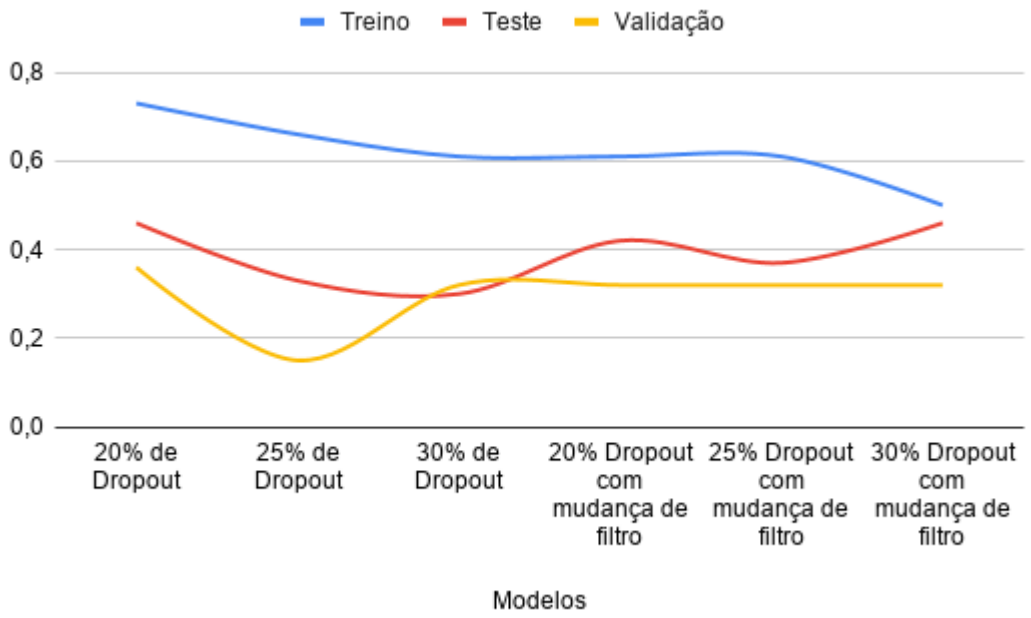
Tabela 01. **Apresentação dos cenários A e B considerando a mudança nos parâmetros da rede neural híbrida**

Componente do modelo	Parâmetros do cenário A	Parâmetros do cenário B
Conv1D (1ª Camada)	128	64
Conv1D (2ª Camada)	256	128
LSTM	128	64
Densa (Dense)	64	64
Dropout (Testados)	20%, 25%, 30%	20%, 25%, 30%

Fonte: Dados da pesquisa

Com base na mudança desses parâmetros, apresenta-se abaixo o **gráfico 1** com os resultados e a análise.

Gráfico 01. **Resultados obtidos do treino, do teste e da validação nos modelos de rede neural híbrida**



Fonte: Dados da pesquisa

Para os modelos do Cenário A, utilizou-se uma estrutura de parâmetros mais robustos com 128 e 256 filtros convolucionais e 128 de unidades LSTM. Essas redes demonstraram maior capacidade de memorização dos dados de treino, atingindo a acurácia máxima observada de aproximadamente 73% com 20% de Dropout. No entanto, o aumento do Dropout para 25% e 30% reduziu progressivamente o desempenho de treino e teste. Destaca-se uma instabilidade na validação quando a taxa de 25% de Dropout foi aplicada e a acurácia sofreu uma queda abrupta para cerca de 15%, sugerindo que o excesso de parâmetros associado a uma taxa intermediária de regularização prejudicou a convergência local do modelo.

No cenário B, em que se fez a redução da complexidade estrutural com 64 e 128 filtros e com 64 unidades LSTM, observou-se uma mudança significativa na dinâmica de aprendizado. O comportamento da curva de validação (linha amarela) tornou-se notavelmente estável, fixando-se de forma linear na faixa dos 32% independentemente da taxa de Dropout aplicada. Isso evidencia que a redução da capacidade da rede atenuou a variância extrema observada no cenário anterior.

Depois da escolha do modelo mais adequado ao trabalho com base nas informações acima, avaliou-se a dinâmica do modelo híbrido com base na convergência e na estabilidade a longo prazo, realizando um ensaio variando o número de épocas de treino de 50 até 200. Os resultados quantitativos de acurácia para os conjuntos de treino, teste e validação encontram-se sistematizados na **tabela 2**, e a sua correspondente evolução temporal está ilustrada no **gráfico 2**.

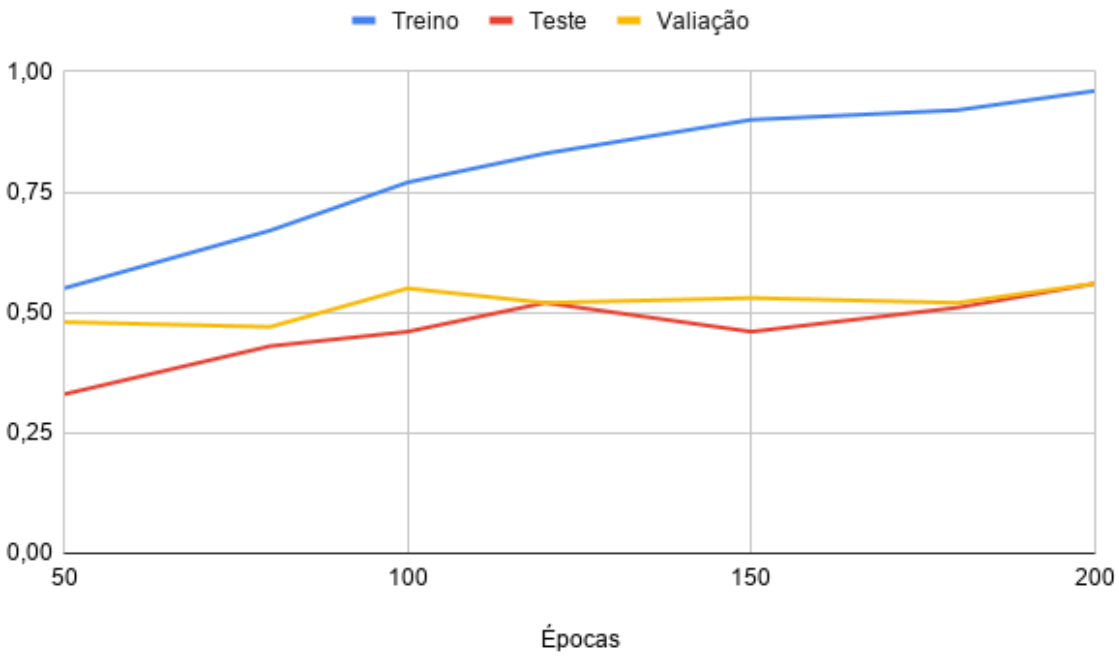
Tabela 02. **Análise da acurácia no comportamento do modelo com 30% de Dropout e com as mudanças nos valores dos filtros nas épocas de treinamento**

Épocas	Acurácia		
	Treino	Teste	Validação
50	0,55	0,33	0,48
80	0,67	0,43	0,47
100	0,77	0,46	0,55
120	0,83	0,52	0,52
150	0,90	0,46	0,53
180	0,92	0,51	0,52
200	0,96	0,56	0,56

Fonte: Dados da pesquisa.

A análise do gráfico mostra que há um crescimento acentuado da acurácia do treino, que inicia em 55% (50 épocas) e atinge o limiar de 96% ao fim de 200 épocas. Este comportamento evidencia a elevada capacidade adaptativa da arquitetura em mapear e memorizar os padrões intrínsecos dos dados de treino. No entanto, ao observar comparativamente os conjuntos de teste e de validação, constata-se que existe uma estagnação no desempenho a partir da época 100. Veja-se que, entre as épocas 120 e 200, a acurácia de validação se estabilizou num patamar entre 52% e 56%. Como consequência disso, o gráfico mostra um hiato de generalização (*generalization gap*), de forma que a diferença entre a acurácia de treino e as métricas de validação e de teste se expandiu, passando de 7% às 50 épocas para 40% às 200 épocas.

Gráfico 02. **Ilustra os valores no comportamento do modelo com 30% de Dropout e com as mudanças no valores dos filtros nas épocas de treinamento**



Fonte: Dados da pesquisa.

Outro ponto que se deve destacar é o fato de que este cenário se caracteriza como um quadro de *overfitting*. À medida que o treino avança acima de 100 épocas, a rede deixa de extrair as características generalizáveis dos áudios e passa a ajustar-se ao ruído específico do conjunto de treino. Embora o pico máximo de validação e teste ocorra às 200 épocas (56%), o custo computacional e o risco de perda na qualidade da classificação tornam os modelos gerados após a época 120 menos viáveis para aplicação prática em cenários reais.

5. Conclusão

Quando se observa o ponto de destaque do experimento que ocorreu no cenário B com 30% de Dropout, vê-se que a acurácia de treinamento reduziu para 50%, enquanto a acurácia de teste alcançou seu segundo maior pico histórico (46%). Consequentemente, o hiato de generalização (*generalization gap*) entre o treino e o teste foi reduzido para apenas 4%. Este comportamento comprova a eficácia da combinação de uma arquitetura mais enxuta com uma maior regularização, resultando no modelo mais robusto, menos propenso ao *overfitting* e com melhor potencial de generalização entre os modelos testados.

É necessário um olhar atento em relação aos dados, pois mesmo apresentando um potencial positivo, os dados de acurácia não são satisfatórios, mas a proposta híbrida apresentou um potencial, pois não apenas memorizou os dados do treino, mas conseguiu classificar, mesmo que ainda não possa ser usado com segurança. Portanto, os resultados do estudo apresentaram um caminho promissor na classificação de sentenças faladas, principalmente por utilizar parâmetros equilibrados no treino e teste do sistema, abrindo possibilidades para novos ajustes e abordagens e melhorias.

Referências

ALMEIDA, Rodrigo José S. et al. Aprendizado de máquina no apoio à transcrição e classificação da fala gaguejada: uma revisão sistemática da literatura. *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, p. 400-411, 2024.

AZEVEDO, Gabriel Almeida et al. *Análise acústica da fala para auxílio à detecção e à classificação de distúrbios da voz*. 2025.

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.

CARVALHO, Cid Ivan C.; OLIVEIRA, Francisca Ticiany B. L.; SILVA, Vitória Maria A.. *Modelo de Classificação Automática de Frases Faladas com Abordagem em Redes Neurais Convolucionais*. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 16. , 2025, Fortaleza/CE. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 519-525. DOI: https://doi.org/10.5753/stil.2025.37852.

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.

DESHPANDE, Kedar et al. A Time-Distributed CNN-LSTM with Attention Model for Speech Based Emotion Recognition. In: *Proceedings of the 2024 7th International Conference on Digital Medicine and Image Processing*. 2024. p. 67-71.

DIJKSTRA, Bauke Alfredo. *Reconhecimento de fonemas utilizando redes neurais convolucionais para transcrição fonética automática*. 2021. Dissertação de Mestrado. Universidade Tecnológica Federal do Paraná.

FACELI, Katti et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. 2011.

KENEDY, Eduardo. O status tipológico das construções de tópico no Português Brasileiro: uma abordagem experimental. *Revista da ABRALIN*, 2014.

LADD, D. Robert; COUTO, Tradução de Letícia Rebollo; SEARA, Izabel Christine. O que é prosódia. *Working Papers em Linguística*, v. 20, n. 1, p. 8-46, 2019.

LI, Charles N.; THOMPSON, Sandra A. Subject and topic: A new typology of language. In: *Subject and topic*. Routledge, 1976. p. 457-489.

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.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

MOURA, Gustavo Bennemann de. *Redes neurais recorrentes para a classificação de estruturas retóricas.* 2018.

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.

SCARPA, E. M (Org). *Estudos de prosódia*. Campinas: Unicamp, 1999.

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.

Wagner, A. (2008). *Automatic labeling of prosody, ITRW on Experimental Linguistics*, ExLing 2008, 25-27 August 2008, Athens, Greece. Disponível em: https://www.isca-archive.org/exling_2008/wagner08_exling.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, IEEE Transactions on Speech and Audio Processing*, 2(4), pages 469–481.