XIII Seminário de Extensão e Inovação
XXVIII Seminário de Iniciação Científica e Tecnológica da UTFPR
Ciência e Tecnologia na era da Inteligência Artificial: Desdobramentos no Ensino Pesquisa e Extensão
20 a 23 de novembro de 2023 - *Campus Ponta Grossa, PR*
SEI-SICITE 2023

Reconhecedor de fonemas sustentado usando redes neurais convolucionadas

Convolutional neural network-based phoneme recognizer

Roberto Luiz Vieira Filho[1], Marcelo de Oliveira Rosa[2]

RESUMO

Este documento detalha um estudo de iniciação científica conduzido por Roberto Luiz Vieira Filho, um estudante de Engenharia da Computação na Universidade Tecnológica Federal do Paraná (UTFPR) em Curitiba. O foco desta pesquisa é a análise do reconhecimento de fonemas com o emprego de redes neurais convolucionais. Foram realizadas três abordagens distintas para a identificação desses fonemas. Na primeira abordagem, os fonemas foram categorizados em seis rótulos, incluindo as vogais (/a/, /e/, /i/, /o/, /u/) e fonemas consonantais. Neste caso, o modelo atingiu uma taxa de precisão de treinamento de 33,60% e uma precisão de teste de 25,18%. Na segunda abordagem, a categorização se limitou às vogais (/a/, /e/, /i/, /o/, /u/), resultando em cinco rótulos. Os resultados revelaram uma precisão de treinamento de 34,64% e uma precisão de teste de 30,63%. Por fim, a terceira abordagem envolveu a divisão dos fonemas em dois rótulos distintos, um para fonemas consonantais e outro para fonemas vocálicos. Nessa configuração, o modelo alcançou uma precisão de treinamento de 64,99% e uma precisão de teste de 61,17% .
PALAVRAS-CHAVE: fonemas; redes neurais convolucionais.

ABSTRACT

This document details a scientific initiation study conducted by Roberto Luiz Vieira Filho, a Computer Engineering student at the Federal Technological University of Paraná (UTFPR) in Curitiba. The focus of this research is the analysis of phoneme recognition using convolutional neural networks. Three different approaches were used to identify these phonemes. In the first approach, the phonemes were categorized into six labels, including vowels (/a/, /e/, /i/, /o/, /u/) and consonant phonemes. In this case, the model achieved a training accuracy rate of 33.60% and a test accuracy of 25.18%. In the second approach, categorization was limited to vowels (/a/, /e/, /i/, /o/, /u/), resulting in five labels. The results revealed a training accuracy of 34.64% and a test accuracy of 30.63%. Finally, the third approach involved splitting the phonemes into two separate labels, one for consonant phonemes and one for vowel phonemes. In this configuration, the model achieved a training accuracy of 64.99% and a test accuracy of 61.17%.
KEYWORDS: convolutional neural networks; phonemes.

INTRODUÇÃO

Este documento descreve a Iniciação científica (IC) do aluno Roberto Luiz Vieira Filho, estudante de Engenharia da Computação na UTFPR, orientado pelo Professor Marcelo de Oliveira Rosa do DAELT-CT. A IC tem como objeto de estudo o reconhecimento de sons (fonemas) vocálicos provenientes de uma base de dados, por meio de redes neurais. As seções seguintes abordarão sobre os métodos utilizados, os resultados e conclusões do trabalho.

[1] Bolsista do(a) PIBIC/UTFPR. Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil. E-mail: robertoluiz@alunos.utfpr.edu.br. ID Lattes: 8388439178913576.
[2] Docente de Engenharia de Controle e Automação (DAELT). Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil. E-mail: mrosa@utfpr.edu.br. ID Lattes: 0897919842779594.

MATERIAIS E MÉTODOS

A base de dados utilizada neste trabalho é a *TIMIT Acoustic-Phonetic Continoous Speech Corpus-Nist Speech* (GAROFOLO et al., 1993), a qual consiste em várias frases faladas por diferentes locutores, com tons, intensidades, ritmos, entonações e pronúncias diferentes. No conjunto de dados utilizado, há um total de 6.095 palavras da língua inglesa, representadas por um conjunto de 63 fonemas, com esse total abrangendo tanto o conjunto de treinamento quanto o de teste. Cada frase possui 4 arquivos: o som é arquivado em um *Waveform Audio File Format* (WAV); os fonemas de cada frase são guardados em um arquivo PHN; as frases e suas durações são guardadas em um arquivo TXT; as palavras separadas por intervalo de tempo em que são faladas são armazenadas em arquivos WRD.

PRÉ-PROCESSAMENTO DO ÁUDIO

Primeiramente, é necessário que o sinal de áudio seja pré-processado para ser aplicado à rede neural, que será descrita na subseção seguinte. Para isso, cada áudio é carregado utilizando a função *load* da biblioteca Librosa (MCFEE; SUÁREZ, 2022), que transforma o arquivo de som em um vetor de números reais. Porém, esse vetor possui um intervalo após o termino da frase, em que não existem palavras, onde está marcado pelo fonema /h#/, logo é interessante utilizar das informações que existem no arquivo TXT que possui a informação de onde é o último som para treino do vetor, e retira o resto.

O processo envolve dividir um arquivo de áudio em partes menores, cada uma com 200 milissegundos de duração, o que equivale a um pequeno fragmento de som. Em seguida, essas partes são deslocadas em intervalos de 50 milissegundos. Esse método permite analisar o som em detalhes, capturando diferentes aspectos ao longo do tempo. Todas as janelas recebem o seguinte tratamento: o tamanho do vetor janela é comparado com a próxima potência de 2, para serem adicionados 0s (zeros) até que seu tamanho sejam iguais. Essa abordagem faz com que ocorra um fenômeno de suavização da curva de densidade espectral de potência.

Além da segmentação do áudio em vetores menores, outro ponto crucial envolve a atribuição de rótulos a cada segmento. Esse processo é guiado pelo arquivo PHN correspondente à gravação atual. Existem duas possibilidades: em um cenário, a janela de segmentação está ligada a apenas um rótulo, logo atribui apenas esse rótulo, e no outro, a janela abrange mais de um rótulo, onde é necessário que copiemos essa janela na mesma quantidade de janelas que ela abrange, atribuindo cada rótulo, treinando assim a mesma janela para diferentes rótulos.

Após realizar esse processo de segmentação, os vetores são transformados em mel-espectrogramas, construídos da seguinte forma:

1. É calculada a transformada de Fourier em tempo curto (STFT), pois para encontrar uma representação em tempo-frequência, o sinal é dividido em segmentos de curta duração (ou janelas), dentro das quais o sinal é considerado aproximadamente estacionário (BALENA, 2019)
2. É calculado o valor absoluto da transformada, pois os valores desse vetor resultante são

números complexos (com componentes de amplitude e fase) - como o valor de fase pouco interessa na a formação de espectrogramas, restringe-se o uso apenas para a componente de amplitude. Aqui já temos o espectrograma em si;

3. A amplitude é transformada em decibéis, para obter uma escala logarítmica, visto que para frequências mais altas, existem não muitas informações úteis;

4. Por fim, é aplicada uma escala de conversão mel para gerar o mel-espectrograma, armazenado para a futura utilização em treinamento e teste das redes neurais.

Cada áudio no *dataset* possui, como mencionado anteriormente, um arquivo PHN que divide o áudio em fonemas. No total, existem 63 fonemas diferentes e como o objetivo dessa pesquisa é reconhecimento de fonemas vocálicos, foi dividido em grupos os fonemas que possuem peculiaridades fonéticas em comum e foi atribuído um valor numérico a cada grupo. Essa abordagem foi utilizada para transformar em uma representação mais manipulável dos fonemas para treino e teste de redes neurais.

- A: recebem o valor de 0 e representam os fonemas ’ae’, ’ao’, ’aa’, ’ah’;
- E: recebem o valor de 1 e representam os fonemas ’eh’, ’ey’, ’iy’, ’er’;
- I: recebem o valor de 2 e representam os fonemas ’ow’, ’oy’, ’aw’, ’uw’,’axr’;
- O: recebem o valor de 3 e representam os fonemas ’ih’, ’ay’,’ix’;
- U: recebem o valor de 4 e representam os fonemas ’uh’, ’ux’,’uw’;
- Sem som: recebem o valor de 6 e representa o fonema ’h#’;
- Consoantes: recebem o valor de 5 e representam os demais fonemas.

Com finalidade de comparação, foi também criada uma classificação para dois rótulos, vogais e consoantes, na qual os fonemas dos correspondentes às vogais recebem o valor 1 e os demais fonemas recebem o valor de 0.

REDE NEURAL

No processo de reconhecimento e classificação dos fonemas, foram empregadas três redes neurais distintas. Cada uma delas apresenta uma estrutura idêntica, com variações identificadas na camada de saída. Essas divergências compreendem a configuração do número de neurônios de saída, bem como a seleção de diferentes funções de ativação. Vale ressaltar que, nesse cenário, optou-se pela função "Sigmoid"para viabilizar a classificação binária, ao passo que a função "Softmax"foi adotada para possibilitar a classificação multiclasse. Para realizar o reconhecimento e classificação dos fonemas, foram utilizadas 3 redes, que possuem duas funções e possuem apenas a camada de saída diferentes. Foi utilizada a biblioteca Keras (CHOLLET; CONTRIBUTORS, 2022). Elas possuem o seguinte layout: camada inicial Conv1D com 32 neurônios e tamanho de kernel de 3; camada

MaxPooling1D com o tamanho de pool de 2; camada intermediária Conv1D com 64 neurônios e tamanho de kernel 3; camada MaxPooling1D com o tamanho de pool de 2; camada Densa com 32 neurônios (para os testes 1 e 2)/ 64 neurônios (para o teste 3).

Como mencionado anteriormente, a configuração da camada de saída varia de acordo com a natureza do teste:

1. Classificar entre vogais (específicas) e consoantes (gerais) - camada de saída com 6 neurônios e ativação softmax;
2. Classificar entre as vogais - camada de saída com 5 neurônios e ativação softmax;
3. Classificar entre vogais (gerais) e consoantes (gerais) - camada de saída com 1 neurônio (classificação binária) e ativação sigmoide;

RESULTADOS

É importante destacar que foram empregadas 20 épocas de treinamento para cada uma das redes utilizadas. Esse valor foi escolhido para que a rede fosse mais treinada, porém, evitando o problema overfitting, onde a rede é treinada bem o suficiente para informações conhecidas, porém tem um desempenho ruim com novos dados. Além disso, devido às limitações do poder computacional, não foi possível utilizar todas as amostras disponíveis em cada treinamento e teste, sendo assim, as amostras que seriam usadas em cada rede eram escolhidas aleatoriamente para ter uma melhor distribuição de rótulos. Sendo assim, aqui estão os resultados:

CLASSIFICAÇÃO ENTRE VOGAIS (ESPECÍFICAS) E CONSOANTES (GERAIS):

Para o primeiro treino, a rede identifica os fonemas vocálicos como os rótulos /a/, /e/, /i/, /o/, /u/ e fonemas consonantais, como sendo os demais. Os conjuntos de fonemas usados no treinamento e teste foram balanceados, ou seja, foram usadas as mesmas quantidades de amostras de fonemas consonantais e vocálicos, com exceção do fonema /u/, que possui menor quantidade de amostras no *dataset* da pesquisa. Na fase de treino, foi obtido um máximo de acerto de $33{,}60\%$ após 20 épocas de treinamento, com uma taxa de crescimento de $0{,}50\%$/época. Na fase de testes, foi obtido um valor de $25{,}18\%$ de acerto na predição.

Rótulo/Fonema	Percentual de acerto do modelo (%)	Quantidade
/a/	20,09	10000
/e/	23,25	10000
/i/	40,28	10000
/o/	18,81	10000
/u/	6,81	4299
Consoante	28,48	10000

Tabela 1 – Tabela de porcentagem de acerto de cada vogal e consoante na fase de testes

Esses valores abaixo da média, com exceção do rótulo /i/, podem ser explicados pelo fato de se atribuir diversos fonemas para 6 rótulos (ou comprimir diversos fonemas em seis), mesmo com

os fonemas podendo não possuir algumas características em comum. Não foram utilizadas muitas amostras, tanto para treino quanto para teste, por conta do poder computacional limitado. Na fase de treino foram utilizadas 159637 amostras, de 842386 disponíveis, enquanto na validação foram utilizadas 54299 amostras de 303968 disponíveis.

CLASSIFICAÇÃO ENTRE VOGAIS:

O segundo treino foi realizado com o intuito de descobrir se as vogais estavam sendo identificadas e a porcentagem de acerto do modelo por fonema vocálico. Foi utilizado um conjunto no qual as vogais /a/, /e/, /i/, /o/ têm a mesma quantidade de amostras, enquanto o /u/ possui menos, por conta da escassez de deste no *dataset* - os demais fonemas foram descartados. No conjunto de treino foi obtido $34{,}64\%$ em 5 épocas. É possível observar também que a precisão do treino não foi muito significativa, tendo uma média de crescimento de $0{,}41\%$/épocas. No conjunto de teste foi obtido um total de $30{,}63\%$ de acerto na predição - a tabela 2 detalha melhor este resultado.

Rótulo/Fonema	Percentual de acerto do modelo (%)	Quantidade
/a/	39,54	10000
/e/	25,14	10000
/i/	32,73	10000
/o/	29,02	10000
/u/	16,27	4299

Tabela 2 – Tabela de porcentagem de acerto de cada vogal na fase de testes

Como é possível observar, todos os fonemas - com exceção de /u/ - apresentaram uma precisão em torno de 30%, um valor bem mais alto do que o encontrado para /u/, por conta do desbalanceamento de amostras (como comentado anteriormente), fazendo com que o modelo não reconheça muito bem amostras com características desse fonema. O valor médio de acerto não é muito significativo e existem algumas causas, como o fato de estarmos agrupando vários fonemas para cada rótulo/fonema e não necessariamente possuem muitas características em comum. Também deve-se considerar que não foram utilizadas todas as amostras, por conta do poder computacional limitado, assim não treinando muito bem os parâmetros da rede neural. Na fase de treino foram utilizadas 163248 amostras de 288369 amostras, enquanto no teste foram utilizadas 44299 amostras de 104605 amostras disponíveis.

CLASSIFICAÇÃO ENTRE VOGAIS (GERAIS) E CONSOANTES (GERAIS):

Por fim, foi realizado uma classificação binária para verificar se o modelo era capaz de diferenciar as vogais de consoantes. Foi utilizado um conjunto balanceado de amostras $50/50$. No conjunto de teste foi obtido um total de $61{,}17\%$, no qual as vogais produziram um melhor resultado, com $63{,}28\%$ e as consoantes $53{,}11\%$ de precisão. Tal resultado pode ser explicado pela diferença da quantidade de fonemas de consoantes e vogais, visto que existem bem mais fonemas consonantais nesse modelo, resultando em um grupo de consoantes com pouquíssimos atributos em comum. Vale ressaltar que foram usadas 20 épocas, para evitar o problema de *overfitting*, produzindo uma

precisão do treino de $64{,}99\%$, com crescimento de $0{,}38\%$. Na fase de treino foram utilizadas 152822

Vogal ou Consoante	Porcentagem de acerto do modelo (%)	Quantidade
Vogal	63,28	16000
Consoante	53,11	16000

Tabela 3 – Tabela de porcentagem de acerto de cada vogal na fase de testes.

amostras, de 842386 disponíveis, enquanto na validação foram utilizadas 32000 amostras de 303968 disponíveis.

CONCLUSÕES

O modelo de classificação de fonemas utilizado foi razoavelmente efetivo para classificar se um som é uma vogal, sem conseguir efetivamente discriminá-la. Seria interessante para trabalhos futuros empregar equipamentos com maior capacidade computacional e um *dataset* maior para que esses modelos sejam capazes de obter resultados mais significativos.

Agradecimentos

O presente trabalho foi realizado com apoio da UTFPR — Brasil (Edital UTFPR/PROPPG n.º 02/2022 — PIBIC)

Conflito de interesse

Declaramos que não há conflito de interesse.

REFERÊNCIAS

BALENA, Eider. **ESTUDO E IMPLEMENTAÇÃO DE ALGORITMOS PARA REPRESENTAÇÃO DE SINAIS EM DISTRIBUIÇÕES TEMPO-FREQUÊNCIA**. [S.l.: s.n.], 2019. Trabalho de Conclusão de Curso. Disponível em: .

CHOLLET, François; CONTRIBUTORS, Keras. **Keras Documentation**. 2022. Disponível em: . Acesso em: 15 out. 2022.

GAROFOLO, John S. et al. **TIMIT Acoustic-Phonetic Continuous Speech Corpus**. [S.l.: s.n.], 1993. Web Download. Philadelphia: Linguistic Data Consortium. LDC Catalog No.: LDC93S1, ISBN: 1-58563-019-5, ISLRN: 664-033-662-630-6, DOI: https://doi.org/10.35111/17gk-bn40.

MCFEE, Brian; SUÁREZ, Juan Luis. **Librosa Documentation**. 2022. Disponível em: . Acesso em: 15 out. 2022.