
**UNIVERSIDADE FEDERAL DO PARÁ**
**INSTITUTO DE TECNOLOGIA - ITEC**
**PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELTÉTRICA**

**RODRIGO GOMES DUTRA**

# TRANSFORMER NETWORKS APLICADA A RECONHECIMENTO AUTOMATICO DE VOZ

**Belém**
**2021**

# 1 INTRODUÇÃO

Este trabalho busca demonstrar o uso preliminar de redes neurais do tipo transformer networks Vaswani et al. (2017) aplicada na tarefa de reconhecimento automático de voz. Nesse sentido a tarefa consiste em classificar palavras isoladas retiradas do banco de dados speach commands Warden (2018), para isso foi utilizado o programa spock Klautau (2021),para pre-processar os sinais de voz em formato WAV, extraindo assim features acústicas utilizando a técnica de Mel-frequency cepstral coefficients (MFCC). Após a extração das features acústicas o modelo então é treinado para a classificação da classe da palavra dado a entrada de features acústica fornecida

Para a tarefa em questão de reconhecimento automático de voz, atualmente as redes neurais do tipo recorrente (RNNS) e long short term memory (LSTM), representam o estado da arte Zeng et al. (2021), juntamente com redes convolucionais. Entretanto, redes neurais do tipo transformers ultrapassam redes neurais do tipo recorrentes em aplicações de processamento de linguagem natural[] e predição de séries temporais, dessa forma apresentam grande potencial para alcançar resultados promissores na área de reconhecimento automatico de voz.

A fim de comparar o resultado do modelo, o programa spock utiliza a técnica de hmm, assim esta será utilizado como baseline. Ambas as tecnicas serão treinadas para reconhecer as seguintes palavras: dog, cat, house, happy e zero. Para a composição do dataset de treino foi escolhido cem audios em formato wav para cada palavra e para teste serão utilizados vinte e cinco audios de cada palavra.

# 2 DESENVOLVIMENTO

## 2.1 Pre-processamento de dados

Reconhecimento automático de voz de palavras isoladas é uma tarefa de aprendizado supervisionado, de forma que o audio é a entrada do modelo e a saída do modelo é a classe da palavra correspondente ao audio de entrada. Dessa forma a fim de reduzir o ruído do sinal de entrada e extrair as features acústicas do sinal é aplicado a técnica de MFCC.

Assim o esquemático do MFCC é como segue:

- Janelamento do sinal
- Transformada de fourier discreta (DFT)
- Mel-filter Bank
- Conversão para escala logarítmica
- Transformada inversa

A etapa de janelamento do sinal consiste em retirar segmentos do sinal, no caso da implementação de 240ms, para posteriormente ser processado e retirado as features acústicas de cada segmento.

A partir desse ponto o sinal então passa pela transformada discreta de fourier, dado que esse sinal de entrada é um audio em formato WAV e trata-se de um sinal digital. Dessa forma é obtido o sinal no domínio da frequência, onde é este pode ser analisado de maneira mais simples.

O Mel-filter bank consiste em um método de imitar a percepção humana do som para as maquinas. O ouvido humano tem uma percepção diferente para diferentes frequências, ao passo que nas maquinas as frequências tem a mesma resolução, então são percebidas da mesma forma. Dessa forma esse método realiza o mapeamento abaixo:

$$mel(f) = 1127ln(1 + \frac{1}{700}) \tag{2.1}$$

Após esse passo é realizado a conversão para escala logarítmica e então é realizado a transformada de fourrier inversa para o dominio do tempo. Nesse passo o algoritimo seleciona os primeiros 12 coeficientes do sinal antes de realizar a transformada inversa. Juntamente com os 12 coeficientes, também é retirado a energia do sinal como uma feature. Juntamente com as 13 features retiradas também é realizado a primeira e a segunda derivada dessas 13 features iniciais, gerando mais 26 features, e o total de 39 features por segmento do sinal inicial.

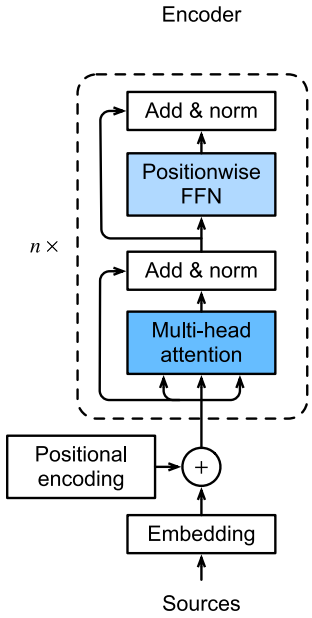
## 2.2 Transformer networks

As transformer networks foram construidas pelo autor Vaswani et al. (2017) para processar dados sequenciais e potencialmente ultrapassar a performance das redes neurais do tipo LSTM para a tarefa de processamento de linguagem natural. Após essa aplicação muitas outras surgiram fora do campo de processamento de linguagem natural, como em séries temporais Wu et al. (2020) Zhou et al. (2021) e até mesmo para classificação de imagens Chen, Fan e Panda (2021). Em todas essas aplicações o modelo baseado em transformer networks consegue ultrapassar os modelos estado da arte tanto em performance quanto em custo computacional, além do fato de possibilitar processamento paralelo entre várias maquinas, tarefa a qual não pode ser efetuada quando utilizado modelos baseados em redes recorrentes.

No contexto de reconhecimento automático de voz, a rede do tipo transformers também foi capaz de atingir performance maior que os modelos considerados estado da arte Lu et al. (2020). Dessa forma é um modelo interessante para se avaliar nesse contexto, de forma prática.

Portanto, foi utilizado o somente o encoder da arquitetura encoder decoder da rede transformer, como mostrado abaixo:

**Figura 1 – Estrutura do transformer encoder**




**Fonte: O Autor (2021)**

O encoder do transformer utiliza como entrada uma matriz de três dimensões: batch, passo de tempo, features. No caso desse trabalho essa estrutura irá ficar da forma batch, segmento, features acústicas.

Diferentemente de redes neurais recorrentes, como a LSTM, a transformer network utiliza do positional encoding para adicionar a noção de ordem temporal no modelo. O positional encoding faz isso por meio de senos e cossenos de diferentes frequencias, inserindo assim a noção de sequencia na matriz de entrada

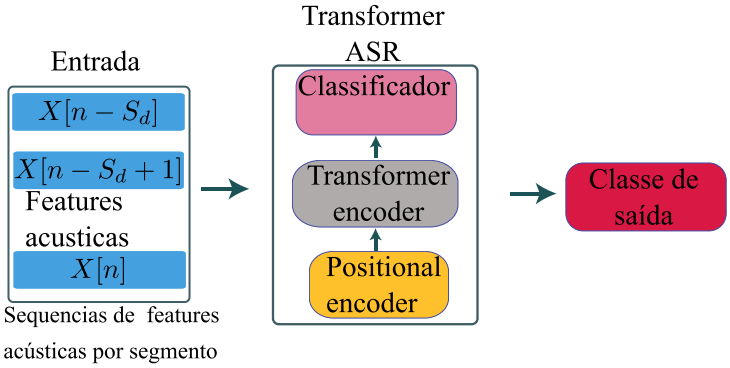
$$PE_{pos,2i} = sin(pos/1000^{2i/d_{model}}), \tag{2.2}$$

$$PE_{pos,2i+1} = cos(pos/1000^{2i/d_{model}}), \tag{2.3}$$

Após o processo do positional encoding a entrada passa pelo processo do encoder, onde o tensor agora irá passar pelo processo das multiplas attention heads, onde é gerado uma matriz de self attention. Essa matriz de self attention é o mecanismo que a rede transformer utiliza para focar nos pontos mais importantes da sequencia de entrada. Depois desse processo é efetuado normalizações e outras operações por camadas do tipo feed forward.

Após o processo do encoder, a informação é passada por uma camada simples do tipo feed forward para realizar a classificação da palavra isolada. A imagem abaixo retrata o fluxo do modelo total de forma simples:

**Figura 2 – Estrutura do modelo transformer**




**Fonte: O Autor (2021)**

Onde, na figura 4 cada vetor $X[n]$ contem as 39 features acústicas, e a entrada total é uma matriz composta por 197 vetores de features acústicas. A implementação completa pode ser acessada em: Dutra (2021).

## 2.3 Baseline

O baseline trata-se do algoritmo que utiliza de hidden markov models (HMMs) para realizar a classificação, dado as features acústicas de entrada. HMMs são modelos estatísticos que podem ser utilizados na área de aprendizado de máquina, para tarefas de classificação. O objetivo desse modelo é aprender a tarefa como uma cadeia de markov, por observar os estados escondidos.

A fim de implementação, o baseline composto por 5 HMMs, com o número de gaussianas variando de 1 a 5, foi executado pelo programa spock Klautau (2021) em forma de guided user interface (GUI) para o processamento dos arquivos WAV em features acústicas e posteriormente para a tarefa de reconhecimento automático de voz utilizando as HMMS.

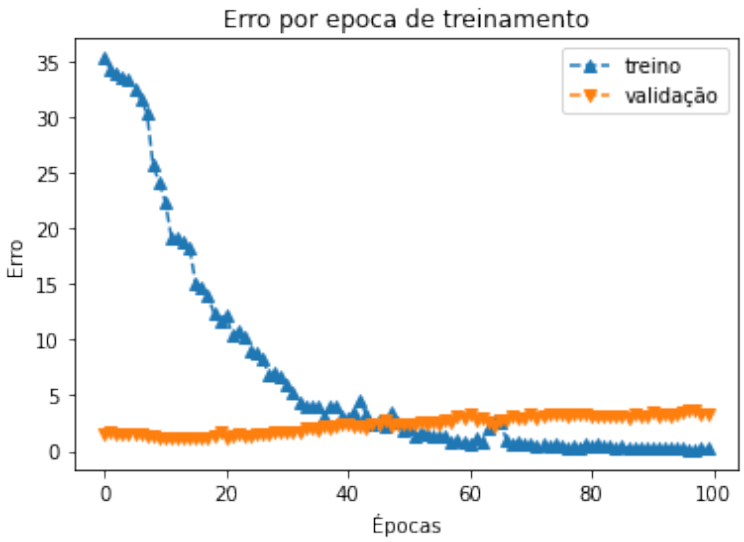
# 3 EXPERIMENTO E RESULTADOS

Os modelos foram postos a prova utilizando o dataset speach commands Warden (2018), segmentando o dataset para conter somente as palavras : dog, cat, house, happy e zero, com 100 amostras sonoras para cada palavra no dataset de treinamento e 25 amostras sonoras de cada palavra para teste.

Após a seleção do dataset, foi executado o programa spock para o pre-processamento dos dados para o modelo baseado em transformer networks e para geração do baseline. Com os audios de entrada ja processados em features acústicas por segmentos, é então iniciado o passo de normalização para utilizar essas features como entrada no modelo transformer.

Depois da normalização, os dados são formatados a fim de entrar uma matriz com dimensões iguais a batch, segmento, features acústicas para o modelo, de forma que o modelo transformer irá traçar a relação entre cada passo de tempo (segmento) utilizando o mecanismo de self attention e a partir disso irá fazer a inferência de classificação. Abaixo segue o erro e a acurácia por época. Onde para a validação do modelo foi utilizado 25 exemplos de audio para cada palavra.

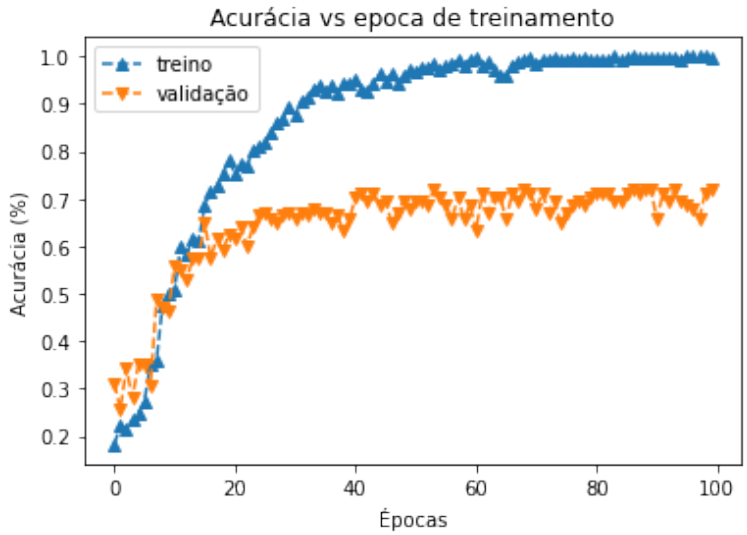
**Figura 3 – Erro por época**




**Fonte: O Autor (2021)**

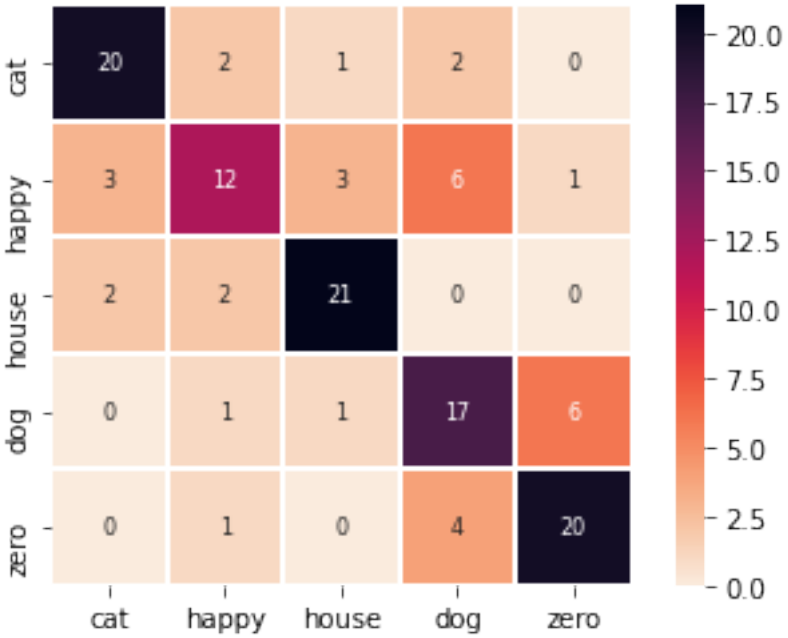
Após o treinamento, o modelo pode realizar a classificação com 72% de acurácia total, demonstrando valores diferentes de acertividade de acordo com a palavra. Como mostrado na matriz de confusão na figura 5.

**Figura 4 – Acurácia por época**




**Fonte: O Autor (2021)**

**Figura 5 – Matriz de confusão**




**Fonte: O Autor (2021)**

Agora comparando-se com o baseline composto por HMMs, temos na tabela 1, que o modelo composto pela rede neural do tipo transformers foi capaz de ultrapassar somente o modelo baseado em HMM que é formado utilizando somente uma gaussiana. Outro fato interessante é que ao aumentar a complexidade dos modelos HMM estes começam a perder performance, o que indica que os modelos podem ter caído no estado de overfitting.

**Quadro 1 – Comparação dos modelos**

| **Modelo** | ***Acurácia total*** |
|---|---|
| Transformer network | 72.0% |
| HMMs com 1 gaussianas | 66.4% |
| HMMs com 2 gaussianas | 72.8% |
| HMMs com 3 gaussianas | 79.2% |
| HMMs com 4 gaussianas | 78.4% |
| HMMs com 5 gaussianas | 76.8% |

# 4 CONCLUSÃO

De posse dos resultados é possível notar uma quebra de expectativa entre a escolha de um modelo que está sendo altamente utilizado para vários problemas com dados sequenciais e um modelo consolidado estatístico probabilístico. De modo que o modelo baseado em transformer networks não ultrapassou a performance do melhor modelo baseado em HMMs como o esperado ao ler trabalhos que retratam o estado da arte do campo de reconhecimento automático de voz utilizando palavras isoladas.

Uma razão para tal resultado é pelo fato de redes neurais complexas precisarem de uma base de dado extensa e representativa para poderem generalizar melhor. Dessa forma talvez a escolha de 100 amostras de audio por palavra pode não ter ajudado a rede transformer a alcançar seu potencial máximo. Ao passo disso é importante notar a relevancia de modelos consolidados na literatura, de forma que mesmo com uma base de dados "pequena"para o deep learning os modelos baseados em HMMs puderam demonstrar seu potencial.

# REFERÊNCIAS

CHEN, C.-F.; FAN, Q.; PANDA, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. **arXiv preprint arXiv:2103.14899**, 2021.

DUTRA, R. G. **ASR transformer**. [S.l.]: GitHub, 2021. <https://github.com/rodgdutra/ASR_transformer>.

KLAUTAU, A. **Spock: Easy speech recognition**. [S.l.]: GitHub, 2021. <https://github.com/aldebaro/easy-speech-recognition>.

LU, L. et al. Exploring transformers for large-scale speech recognition. **arXiv preprint arXiv:2005.09684**, 2020.

VASWANI, A. et al. Attention is all you need. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2017. p. 5998–6008.

WARDEN, P. Speech commands: A dataset for limited-vocabulary speech recognition. **arXiv preprint arXiv:1804.03209**, 2018.

WU, N. et al. Deep transformer models for time series forecasting: The influenza prevalence case. **arXiv preprint arXiv:2001.08317**, 2020.

ZENG, J. et al. Semi-supervised training of transformer and causal dilated convolution network with applications to speech topic classification. **Applied Sciences**, Multidisciplinary Digital Publishing Institute, v. 11, n. 12, p. 5712, 2021.

ZHOU, H. et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: **Proceedings of AAAI**. [S.l.: s.n.], 2021.