14° Seminário de Iniciação Científica e Tecnológica

AVALIAÇÃO DE REDES NEURAIS PROFUNDAS NO RECONHECIMENTO DE VOZ

Bruno de Araújo Alves (IC), *Hugo Vinícius Leão e Silva (PQ)

PIBITI
Câmpus Anápolis
** hugo.vinicius@ifg.edu.br*

***Palavras Chave**: Reconhecimento de Voz; Aprendizagem de Máquina; Processamento de Sinais de Áudio*

Introdução

Tecnologias de reconhecimento biométrico são cada vez mais comuns. Dentre as tecnologias que mais estão se tornando importantes na atualidade estão as tecnologias baseadas no reconhecimento de voz. O motivo para isso é constante evolução da área de Aprendizado de Máquina nos últimos anos, principalmente após o advento de técnicas de Aprendizagem profunda. Para tanto, é necessário realizar técnicas de processamento de áudio, além da extração de características do sinal de voz. Levando isso em conta, o objetivo deste trabalho é avaliar diversos métodos de aprendizado de máquina, assim como diferentes extratores de características do sinal de voz, visando avaliar quão eficientes e eficazes podem ser esses métodos de aprendizagem no reconhecimento de pessoas.

Metodologia

Inicialmente buscou-se na internet um conjunto de dados com gravações de vozes de diversas pessoas. Foi encontrado um conjunto com mais 960 amostras de áudio de 96 falantes com diferentes sotaques da língua portuguesa. A partir disso, o projeto iniciou etapa de pré-processamento do conjunto de dados encontrado. Para tanto, realizaram-se o recorte de trechos de silencio no sinal de áudio, a segmentação do áudio em trechos de tamanho fixo, a introdução de ruído branco. Por fim, geraram-se três representações para cada áudio com a aplicação de métodos extratores de característica de áudio. Como métodos extratores, utilizaram-se os métodos mais comuns na literatura: *Mel-Frequency Cepstrum Coefficients* (MFCC), *Linear Predictive Coding* (LPC) e *Linear Predictive Coding Coefficients* (LPCC). Logo após, iniciaram-se o treinamento de algoritmos de aprendizado de máquina, utilizando as representações resultantes da etapa anterior, gerando um modelo para cada combinação de algoritmo e representação. Foram avaliados três algoritmos de aprendizado de máquina, *Support Vector Machines* (SVM), *Multi-Layer Perceptron* (MLP) e *Convolutional Neural Networks* (CNN). Por fim, avaliou-se a taxa de acerto de cada modelo obtido, além de realizar a inferência com áudios fora do conjunto de dados de treinamento.

Resultados e Discussão

Cada um dos algoritmos de aprendizado de máquina, SVM, MLP e CNN, foi avaliado em dois estágios. Primeiramente, avaliou-se a taxa de acerto de cada modelo, sendo possível ter uma estimativa do aprendizado de cada modelo no reconhecimento de voz. É importante que o MFCC apresentou os melhores resultados dentre as diversas representações e as versões utilizadas são as implementadas em duas bibliotecas: *Tensorflow* (TF) e *Python Speech Features* (PSF). Assim, a Tabela 1 apresenta apenas as taxas de acerto considerando o MFCC para cada um dos modelos utilizando áudios do conjunto de dados. É possível ver que em todos os casos os modelos aprenderam bem os dados tendo uma taxa mínima de acerto de 99%. Posteriormente, avaliou-se a taxa de acerto de cada modelo ao realizar a inferência com 32 áudios de dois participantes da criação do conjunto de dados, ressaltando que estes áudios não estão presentes no conjunto de dados original e foram colhidos apenas para esta etapa do projeto. Na Tabela 2, por sua vez, é apresentada a taxa de acerto dos modelos ao ser realizado a inferência, é perceptível que dentre todos os modelos, o SVM, ainda com uma taxa de acerto muito baixa, 46,87%, obteve o melhor resultado, sendo o modelo com pior resultado o CNN, 9,38%.

Tabela 1. Taxa de acerto dos modelos de aprendizado de máquina utilizando dados de teste do conjunto de dados.

Algoritmo	Extrator	Taxa de acerto
CNN	MFCC (TF)	100%
CNN	MFCC (PSF)	99,95%
MLP	MFCC (TF)	99,87%
MLP	MFCC (PSF)	99,87%
SVM	MFCC (TF)	100%
SVM	MFCC (PSF)	100%

Tabela 2. Taxa de acerto dos modelos de aprendizado de máquina utilizando outro conjunto de dados.

Algoritmo	Extrator	Taxa de acerto
CNN	MFCC (TF)	9,38%
CNN	MFCC (PSF)	9,38%
MLP	MFCC (TF)	12,5%
MLP	MFCC (PSF)	12,5%
SVM	MFCC (TF)	46,87%
SVM	MFCC (PSF)	46,87%

Conclusões

Realizou-se a avaliação de alguns métodos extratores de características juntamente a diversos algoritmos de aprendizado de máquina para o exercício de reconhecimento de falante. Apesar de apresentar ótimos resultados até a fase de teste do modelo com dados do próprio conjunto de dados, obteve-se uma baixíssima taxa de acerto ao realizar-se a inferência. Assim, deve ainda realizar vários testes e avaliações desde a escolha de um conjunto de dados para buscar melhores resultados em trabalhos futuros.

WANG, M.; DENG, W. **Deep face recognition: A survey**. arXiv preprint arXiv:1804.06655,2018.

LOGAN, Beth. **Mel frequency cepstral coefficients for music modeling**. In International Symposium on Music Information Retrieval. 2000.