CENTRO UNIVERSITÁRIO DA FEI

ARIEL GRAÇA FERREIRA

ESTUDO COMPARATIVO ENTRE TÉCNICAS DE EXTRAÇÃO DE CARACTERÍSTICAS DE SINAIS ACÚSTICOS

e

influência em aprendizado profundo aplicado na estimativa de faixa etária pela voz

SÃO BERNARDO DO CAMPO

2021

Ariel Graça Ferreira

Estudo Comparativo Entre Técnicas de Extração de Características de Sinais Acústicos

Qualificação apresentada como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica, pelo Programa de Pós-Graduação em Engenharia Elétrica do Centro Universitário da FEI.

Orientador: Prof. Dr. Ivandro Sanches

São Bernardo do Campo

2021

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MFCC	Mel Frequecy Cepstral Coeficients
LPC	Linear PredictionCoeficients
LPCC	Linear Prediction Cepstral Coeficients
PLP	Perceptual Linear Prediction
RASTA	Relative Spectral Processing
LSF	Line Spectral Frequencies
PCA	Principal Component Analysis
GMM	Gaussian Mixture Models
ASR	Automatic Speech Recognition
WER	Word Error Rate

SUMÁRIO

1 INTRODUÇÃO

Este trabalho tem por finalidade estudar diferentes técnicas de extração de características de um sinal acústico a fim de determinar vantagens e desvantagens de cada modelo.

Obter informações de um sinal sonoro é uma etapa muito importante em diversos sistemas desenvolvidos para a indústria musical, a área de biomedicina, setores de produção e serviços variados, segurança, entre outros, portanto, possuir formas de caracterizar o sinal e então manipular as informações desejadas, de forma cada vez mais precisa, e eficiente, é algo de grande relevância para o desenvolvimento científico. Através do sinal da nossa voz é possível obter muitas informações sobre o indivíduo que produz o som, como gênero, emoções e idade, porém, devido a diversos fatores que serão discutidos no decorrer da pesquisa, em geral o sinal se encontra exposto a formas de ruído e/ou distorções, oriundas até mesmo da própria fonte, que dificultam a extração das características acústicas (features), dificultando assim a obtenção de informações e o processamento digital do sinal.

Vejo o processo de caracterização de um sinal sonoro, que será processado digitalmente, como a etapa de obtenção da "matéria prima" para os sistemas que operam com este tipo de processamento. Quanto melhor for a qualidade da matéria prima, melhor será o resultado apresentado pelo sistema, ou seja, melhor o seu desempenho. Esse é o propósito principal desta pesquisa, estudar os algoritmos já difundidos no mercado e no meio científico, trabalhar na implementação de cada um deles para então verificar, de forma prática, o desempenho de um mesmo sistema quando utiliza features extraídas a partir de cada uma das técnicas abordadas.

Com base na revisão bibliográfica realizada até o momento, as técnicas selecionadas são: MFCC, LPC, LPCC, PLP, RASTA e LSF.

O intuito do estudo não é comparar as técnica objetivando definir qual seria a melhor ou pior, mas termos um mapeamento das vantagens e desvantagens dos principais algoritmos, onde tais características sejam validadas sobre as mesmas condições, em um mesmo cenário de utilização. Planejo utilizar um sistema de estimação de faixa etária que utiliza algoritmos clássicos de classificação de dados (GMM e i-vectors), porém também está no plano de trabalho utilizar o mesmo sistema com técnicas mais modernas de classificação, Redes Neurais, e assim coletar mais dados empíricos para as análises e comparações finais deste estudo.

O banco de dados adotado para a realização dos ensaios e simulações, será um banco de gravação de sinais coletados por chamadas telefônicas de 770 pessoas, totalizando 47 horas de gravações de áudio. O banco pertence a *Deutsche Telekom AG Laboratories*, de Berlim na Alemanha.

O sistema de estimação de faixa etária selecionado para ser utilizado neste trabalho foi desenvolvido pelo Prof. Dr. Sanches (2018), o banco de dados também é o mesmo utilizado no

trabalho do professor.

Para continuação das pesquisas e estudos, posteriores ao Mestrado, tenho interesse em expandir as pesquisas em algumas vertentes. Uma delas seria a busca de formas eficientes de otimizar a utilização de banda para transmissão e armazenamento de dados (especialmente dados provenientes de sinais acústicos), me aprofundando assim no estudo da codificação e compressão desse tipo de dado. Vejo esse projeto atual, como a base para o início de uma linha de pesquisa, pois com ele poderei ter um conhecimento sólido, e de baixo nível, que acredito ser de suma importância para estudos futuros.

1.1 Objetivos

- Comparar técnicas de extração de características de um sinal acústico a fim de determinar vantagens e desvantagens de cada algoritmo.

- Fazer a implementação de tais técnicas para utilização junto a um sistema de estimação de faixa etária a fim de medir o desempenho do mesmo e, consequentemente, avaliar a eficácia de cada algoritmo. Nessa primeira parte, o sistema utiliza GMM e i-vector para classificação dos dados.

- Em uma segunda etapa, as mesmas técnicas de extração de features serão implementadas junto a um sistema de estimação de faixa etária, porém dessa vez a classificação dos dados será realizada a partir de uma rede neural. Pretendo assim medir e comparar a influência dos métodos de extração de features em relação aos dois tipos de sistema.

- Tabelar as principais características de cada algoritmo e apontar casos/cenários de uso mais indicados para cada um deles.

- Investigar e avaliar a viabilidade do uso de algoritmos combinados.

2 REVISÃO BIBLIOGRÁFICA

Existem trabalhos acadêmicos que também promovem uma comparação entre diversas técnicas de caracterização de sinais sonoros. Um desses estudos foi publicado por Khan et al. (2019), onde são abordadas as técnicas MFCC, PLP, PCA, LPC, LPCC e RASTA, sendo que o intuito é mapear, para cada um desses modelos, nível de complexidade quanto a implementação, precisão e desempenho utilizando dados ruidosos. No trabalho em questão, os autores utilizam apenas dados teóricos, baseando discussões apenas em cálculos matemáticos, enquanto na pesquisa que proponho haveria uma evolução desse trabalho, trazendo dados práticos para a discussão do tema.

Outro artigo, publicado pelos pesquisadores Alim e Rashid (2018), segue na mesma linha de pesquisa, acrescentando porém o algoritmo LSF, embora os autores também tenha se atido a dados obtidos em análises puramente teóricas.

As duas publicações supra citadas, como mencionei, são dois exemplos de pesquisas já realizadas e que estão na mesma linha do estudo que pretendo seguir durante o programa de Mestrado. Além destes, gostaria de citar ainda outros dois artigos que me chamaram atenção e contribuíram com ideias para algumas pesquisas.

O primeiro seria o artigo do Pellegrini et al. (2014), onde os autores discorrem sobre a queda de desempenho de sistemas de reconhecimento automático de fala (ASR) quando utilizados por idosos, visto que haveria uma certa necessidade de adaptar tais sistemas para este tipo de público devido a eventual descaracterização descaracterização dos sinais de voz decorrente do avanço da idade. Acredito que este seja um tema bem interessante para ser explorado e verificar se a etapa de extração de features ainda pode evoluir de alguma forma para mitigar este problema.

O segundo trabalho seria a publicação de Joshi e Cheeran (2014). Neste, os autores apresentam uma implementação de MFCC com Matlab, porém o grande valor do trabalho está na descrição etapa por etapa de um processo de extração de características de um sinal de áudio, nesse caso utilizando MFCC's como mencionado. Ainda de acordo com Joshi e Cheeran (2014), podemos considerar o LPC, MFCC e o PLP como as três principais técnicas de extração de features. Levando em conta que a publicação desse trabalho ocorreu em 2014, e considerando também outras técnicas mais modernas que foram e estão sendo estudadas em pesquisas mais recentes, acredito que o LPC, MFCC e PLP são realmente técnicas muito importantes e de grande popularidade, tendo sido talvez, de alguma forma, precursoras de outras técnicas que estão sendo utilizadas em sistemas modernos, todavia é notório que esse nicho de estudos têm se expandido e outras técnicas vêm sendo difundidas e ganhando grande destaque no decorrer dos anos.

Gostaria de mencionar aqui também alguns livros que têm servido como base para que

algoritmo

eu possa entender, de forma mais sólida, não apenas a teoria matemática por trás dos algoritimos mais tradicionais como o MFCC e o LPC, mas também têm sido a literatura fundamental para minha compreensão do procesamento digital de sinais acústicos no geral. Seriam eles Gold, Morgan e Ellis (2011) e Rabiner e Schafer (2011).

O livro da Zheng e Casari (2018) tem me auxiliado a compreender a abrangência de toda a área de engenharia de features e a importância do estudo do tema para o segmento de ciência de dados.

Sugiro incluir artigos de estimação de faixa etária e deep learning

3 METODOLOGIA

3.1 Sistema de Estimação de Faixa Etária

Para a execução das simulações e ensaios, será utilizado, inicialmente, o mesmo sistema proposto em Sanches (2018), ou seja, um sistema de estimação automática de faixa etária, que utiliza técnicas de aprendizado de máquina como GMM e i-vector, embora, como já mencionado, o sistema será adaptado para que uma rede neural também seja utilizada como classificador e assim seja possível comparar o desempenho dos dois tipos de modelos com todos as técnicas de caracterização de sinais.

3.2 Base de Dados

Base de dados da *Deutsche Telekom AG Laboratories*.

A base de dados utilizada neste trabalho, a qual também foi utilizada por Sanches (2018), é composta por gravações telefônicas realizadas por 852 indivíduos alemães. Em geral, a maioria dos arquivos de áudio possuem 2 segundos de duração, mas também existem arquivos de 3 e 6 segundos. Para a elaboração desse banco, foram selecionados, aproximadamente, 100 indivíduos por classe (balanceados também por gênero), sendo que a divisão das classes segue a tabela abaixo:

Tabela 1 – Classes

Classe	Idade	Gênero
1	7 - 14	f, m
2	15 - 24	f
3	15 - 24	m
4	25 - 54	f
5	25 - 54	m
6	55 - 80	f
7	55 - 80	m

f: feminino / m: masculino

No total, são 47 horas de gravações de áudio, que serão utilizadas para as etapas de treino, teste e validação do sistema. Os arquivos foram disponibilizados no formato .raw e amostrados em 8KHz. Para que seja mais fácil manipular os arquivos com as bibliotecas Python disponíveis, todos serão convertidos para o formato wav, utilizando um código escrito em linguagem de programação Python. Segue um trecho do script criado para conversão dos arquivos:

Figura 1 – Script de conversão de arquivos.

```
def fileDF(PATH, file_train, file_test):
    # header = None, nao considera a primeira linha do txt como header de tabela
    file_tr = pd.read_table(file_train, delimiter = ' ', header = None)
    file_list_tr = file_tr[0]
    file_tt = pd.read_table(file_test, delimiter = ' ', header = None)
    file_list_tt = file_tt[0]
    classe_tr = pd.read_table(file_train, delimiter = '/', header = None)
    classe_list_tr = classe_tr[2]
    classe_tt = pd.read_table(file_test, delimiter = '/', header = None)
    classe_list_tt = classe_tr[2]

    file_list_tr = file_list_tr.str.replace(r'.raw', '.wav')
    file_list_tt = file_list_tt.str.replace(r'.raw', '.wav')

    train_files = file_list_tr.str.replace(r'.raw', '.wav').str.slice(start=22)
    test_files = file_list_tt.str.replace(r'.raw', '.wav').str.slice(start=22)

    train_full_df = pd.DataFrame(file_list_tr)
    test_full_df = pd.DataFrame(file_list_tt)

    train_df = pd.DataFrame(train_files)
    test_df = pd.DataFrame(test_files)

    train_full_df = train_full_df.rename(columns={0: 'file'})
    test_full_df = test_full_df.rename(columns={0: 'file'})

    train_df = train_df.rename(columns={0: 'file'})
    test_df = test_df.rename(columns={0: 'file'})

    return (train_full_df, test_full_df, train_df, test_df, classe_list_tr, classe_list_tt)
```

Script para conversão de arquivos .raw em .wav.

3.3 Métricas

Os parâmetros que pretendo avaliar sobre cada uma das técnicas abordadas e sobre o sistema utilizado nas simulações:

- Tipo de coeficiente.
- Custo computacional para a etapa de extração de features.
- Robustez quanto a adição de ruídos.
- Robustez quanto a degradações na fonte.
- Precisão do sistema (Recall e F-Measure).

3.4 Implementações

3.4.1 Classificador com Rede Neural (A1)

Utilizando a linguagem de programação Python e a biblioteca Keras (que por sua vez, se origina na biblioteca TensorFlow), está sendo desenvolvido um classificador, com base em aprendizado de máquina, para simulação do sistema de estimação de faixa etária.

Alguns trechos do código, apenas para ilustração:

Figura 2 – Rede Neural 1.

```python
import numpy as np
import matplotlib.pyplot as plt
import tensorflow as tf
import os
import sys
import librosa
# import htk_featio as htk
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from keras.utils.np_utils import to_categorical
from sklearn.preprocessing import StandardScaler
from keras.models import Sequential
from keras.layers import Dense, Dropout  # Activation, Flatten
# from keras.callbacks import EarlyStopping
import time


PATH = '/home/spilborghs/Documents/SER_DB/enterface_database/AUDIO_ONLY/'


def Feat_extract(files):
    file_name = os.path.join(os.path.abspath(PATH+str(files.file)))
    # Generate Mel-frequency cepstral coefficients (MFCCs) from a time series
    X, sample_rate = librosa.load(file_name, res_type='kaiser_fast')
    # Generates a Short-time Fourier transform (STFT) to use in the chroma_stft
    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T,
                    axis=0)
    # Computes spectral contrast
    mel = np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T, axis=0)
    return mfccs, mel


# nada, Bt, epoch = sys.argv

Bt, epoch = 1, 10

# FEAT_PATH = "/home/spilborghs/Documents/mel_filter_banks_features/feat/Experiments/"


Train_file = pd.read_csv("Csv/Train_wav.csv")
Test_file = pd.read_csv("Csv/Test_wav.csv")
Val_file = pd.read_csv("Csv/Valid_wav.csv")
Arq_train = Train_file['Arquive']
Arq_test = Test_file['Arquive']
Arq_val = Val_file['Arquive']
Res_train = Train_file['Results']
Res_test = Test_file['Results']
Res_val = Val_file['Results']
```

Classificador em Python - Rede Neural.

Figura 3 – Rede Neural 2.

```python
X_test = np.array(features_test)
X_train = np.array(features_train)
Y_train = np.array(Res_train)
X_val = np.array(features_val)
Y_val = np.array(Res_val)

lb = LabelEncoder()
Y_train = to_categorical(lb.fit_transform(Y_train))
Y_val = to_categorical(lb.fit_transform(Y_val))

ss = StandardScaler()
X_train = ss.fit_transform(X_train)
X_val = ss.transform(X_val)
X_test = ss.transform(X_test)

model = Sequential()
model.add(Dense(168, input_shape=(168,), activation='relu'))
model.add(Dropout(0.1))
model.add(Dense(84, activation='relu'))
model.add(Dropout(0.25))
model.add(Dense(84, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(6, activation='softmax'))
model.compile(loss='categorical_crossentropy', metrics=['accuracy'],
              optimizer='adam')
early_stop = EarlyStopping(monitor='val_loss', min_delta=0, patience=100,
#                           verbose=1, mode='auto')

N_epochs = int(epoch)
BSize = int(Bt)
ini_train = time.time()
history = model.fit(X_train, Y_train, batch_size=BSize, epochs=N_epochs,
                    validation_data=(X_val, Y_val)),
                    callbacks=[early_stop])
model.save("model_"+str(N_epochs)+'_'+str(BSize))

fim_train = time.time() - ini_train
file = open("../../Dados_NN/librosaTestes.txt", 'a+')
file.write("Teste com "+str(N_epochs)+" epocas e "+str(BSize)+" de batch: \n")
file.write("Tempo de Retirada de caracteristicas: "
           + str((1.0*fim_feat)/len(Res_test))+'\n')
file.write("Tempo de Treino: "+str(fim_train)+'\n')

```

Classificador em Python - Rede Neural.

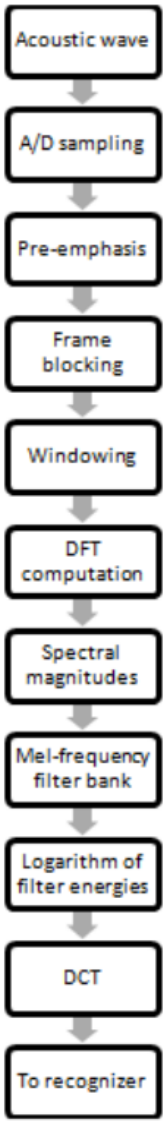
Os primeiros testes com este classificador serão realizados utilizando apenas features extraídas com MFCC.

Para calcular todos os coeficientes, de todos os arquivos de áudio que compõe o banco de dados, o programa criado gasta aproximadamente 60 min.

3.4.2 Implementação MFCC (A2)

Será avaliada a viabilidade de se implementar o algoritmo referente ao MFCC, sem a utilização de bibliotecas prontas. Como visto no tópico anterior, existem diversas formas de se extrair esse tipo de coeficientes de forma direta utilizando ferramentas prontas.

Figura 4 – MFCC.



Fonte: Diagrama de Blocos MFCC - Joshi e Cheeran (2014).

3.4.3 Implementação dos demais algortimos (A3)

Para sequência das simulações e estudos, deve ser implementado os demais algoritmos de extração de features selecionados (LPC, LPCC, PLP, RASTA e GMM) para que todos sejam utilizados em simulações com os dois tipos de classificadores mencionados anteriormente.

3.4.4 Simulações e Ensaios (A4)

Assim que finalizada as implementações dos algoritmos, o plano de trabalho prevê a realização de diversas simulações com os dois tipos de classificadores, utilizado cada uma das técnicas de caracterização de sinal, visando a coleta da maior quantidade de dados possível bem como variando os cenários de forma que seja possível simular potenciais ambientes de utilização e assim medir os efeitos e influência de cada técnica no desempenho dos classificadores.

3.4.5 Discussão dos Dados (A5)

Com os dados em mãos, o próximo passo será analisar detalhadamente todas as informações obtidas, avaliá-las adequadamente e discutir as conclusões com base ~~em toda a~~ na referência teórica pesquisada sobre os temas.

4 CRONOGRAMA

A Tabela 2 apresenta o cronograma de execução das atividades desta proposta.

Tabela 2 – Cronograma das atividades previstas

Etapa	Meses											
	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Revisão bibliográfica	■	■	■	■	■	■	■	■	■	■	■	■
A1		■	■	■								
Banca de qualificação				■								
A2				■								
A3				■	■	■	■					
A4								■	■			
A5										■	■	■

A1: Implementação classificador rede neural e primeiras simulações.

A2: Implementação do algoritmo MFCC.

A3: Implementação dos demais algoritmos.

A4: Execução dos ensaios e simulações para coleta dos dados experimentais.

A5: Elaboração da dissertação, análise e discussão dos resultados.

REFERÊNCIAS

ALIM, S. A.; RASHID, N. K. A. In: ____. *From Natural to Artificial Inteligence - Algorithms and Applications*. [S.l.]: IntechOpen, 2018. cap. Some Commonly Used Speech Feature Extraction Algorithms. Citado na página 8.

GOLD, B.; MORGAN, N.; ELLIS, D. *Speech and udio signal processing: processing and perception of speech and music*. Hoboken, New Jersey, USA: John Wiley & Sons, 2011. Citado na página 9.

JOSHI, S. C.; CHEERAN, D. A. Matlab based feature extraction using mel frequency cepstrum coefficients for automatic speech recognition. *International Journal of Science, Engineering and Technology Research (IJSETR)*, v. 3, n. 6, junho 2014. Citado 2 vezes nas páginas 8 e 13.

KHAN, I. et al. *Robust Feature Extraction Techniques in Speech Recognition: A Comparative Analysis*. Pakistan, 2019. 6 p. Citado na página 8.

PELLEGRINI, T. et al. Speaker age estimation for elderly speech recognition in european portuguese. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 15. *anais*. Singapore, 2014. Citado na página 8.

RABINER, L. R.; SCHAFER, R. W. *Theory and Applications of Digital Speech Processing*. Upper Saddle River, New Jersey, USA: Pearson, 2011. Citado na página 9.

SANCHES, I. *Estimação automática de faixa etária pelo processamento do sinal de voz*. São Bernardo do Campo, 2018. 14 p. Processo FAPESP 2016/18700-7. Citado 2 vezes nas páginas 6 e 10.

ZHENG, A.; CASARI, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, California, USA: O'Reilly, 2018. Citado na página 9.