MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

ABDELHAMID IBN BADIS UNIVERSITY - MOSTAGANEM

Faculty of Exact Sciences and Computer Science

Department of Mathematics and Computer Science

Sector: Computer science

FINAL THESIS

For Obtaining the Master Degree in Computer Science

Option: Artificial Intelligence for the Internet of Things (AI4IoT)

Presented By:

Derdour Amina

THEME:

Artificial Intelligence for Voice-Based Age Estimation

Defended on: June 2, 2024

Board of Examiners:

Dr. Henni Fouad	Supervisor	University of Mostaganem
Pr. Larbi Boubchir	Co-Spervisor	University of Paris-8
Dr. Mohammed Habib Zahmani	President	University of Mostaganem
Dr. Mohammed Elamine Moumene	Examiner	University of Mostaganem

Academic Year: 2023-2024

Résumé

L'estimation de l'âge basée sur la voix est un domaine d'étude émergent avec des applications importantes dans la sécurité biométrique, les soins de santé et les services personnalisés. Notre travail se concentre sur le développement et l'évaluation d'une solution basée sur la mémoire à long terme (LSTM) formée sur l'ensemble de données Common Voice, ciblant spécifiquement la démographie anglophone dans diverses régions.

L'objectif principal de cette étude est de fournir des estimations précises de l'âge à partir de données vocales. Notre modèle extrait les caractéristiques spectrales et les coefficients cepstraux de fréquence de Mel (MFCCs) des échantillons vocaux, en tenant compte du sexe et de l'accent du locuteur pour mieux estimer l'âge. Pour résoudre ce problème, nous avons implémenté notre solution en utilisant Python dans l'environnement Jupyter Notebook, en utilisant des outils tels que Keras pour la création de modèles et Librosa pour le traitement du son. Les résultats sont très encourageants.

Mots-clés : Estimation de l'âge basée sur la voix, apprentissage profond, LSTM, MFCC, Python, Jupyter Notebook

Abstract

Voice-based age estimation is an emerging field of study with significant applications in biometric security, healthcare, and personalized services. Our work focuses on the development and evaluation of a Long Short-Term Memory (LSTM) based solution trained on the Common Voice dataset, specifically targeting English-speaking demographics across various regions.

This study's primary focus is providing accurate age estimates from voice data. Our model extracts spectral features and Mel Frequency Cepstral Coefficients (MFCCs) from voice samples, taking into consideration the gender and accent of the speaker to better estimate the age. To tackle this problem, we implemented our solution using Python in the Jupyter Notebook environment, employing tools such as Keras for model creation and Librosa for sound processing. The results are very encouraging.

Keywords: Voice-based age estimation, deep learning, LSTM, MFCC, Python, Jupyter Notebook

ملخص

تقدير العمر القائم على الصوت هو مجال دراسي ناشئ ذو تطبيقات مهمة في الأمان البيومتري والرعاية الصحية والخدمات الشخصية. يركز عملنا على تطوير وتقييم حل قائم على الذاكرة قصيرة المدى (LSTM) مدرب على مجموعة بيانات Common Voice، ويستهدف على وجه التحديد التركيبة السكانية الناطقة باللغة الإنجليزية عبر مناطق مختلفة.

ينصب التركيز الأساسي لهذه الدراسة على توفير تقديرات عمرية دقيقة من البيانات الصوتية. يستخلص نموذجنا الميزات الطيفية ومعاملات Mel Frequency Cepstral Coefficient (MFCC) من عينات الصوت، مع الأخذ في الاعتبار جنس ولهجة المتحدث لتقدير العمر بشكل أفضل. لمعالجة هذه المشكلة، قمنا بتنفيذ حلنا باستخدام Python في بيئة Jupyter Notebook، باستخدام أدوات مثل Keras لإنشاء النماذج وLibrosa لمعالجة الصوت. النتائج مشجعة للغاية.

كلمات مفتاحية: تقدير العمر القائم على الصوت، التعلم العميق

Dedications

I dedicate this work to my dear family, whose unwavering support, encouragement, and sacrifices have been the foundation of my success.

To my parents, for their endless love, wisdom, and belief in me.

To my sisters, for their constant inspiration and companionship.

To my grandfather, peace to his soul.

To my grandmother, who always supports me with her prayers.

This research is dedicated to you.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Prof. Henni Fouad and Prof. Boubchir Larbi, for their invaluable guidance and advice throughout the stages of writing and implementing this project.

I also extend my sincere thanks to my parents and sisters for their unwavering support and encouragement during this journey. Additionally, I am grateful to my friends, who have been my companions and confidants along the way.

Thank you all.

List of figures

List of tables

List of abbreviations

Abbreviation	Complete Expression	Page
AI	Artificial Intelligence	4
DL	Deep Learning	4
NN	Neural Network	4
DNN	Deep Neural Network	5
ML	Machine Learning	7
IVR	Interactive Voice Response	7
GMM	Gaussian Mixture Model	8
SVM	Support Vector Machine	8
MFCC	Mel-Frequency Cepstral Coefficients	8
PLP	Perceptual Linear Predictive	8
CNN	Convolutional Neural Network	8
LSTM	Long-Short Term Memory	8
MAE	Mean Absolute Error	8
ReLU	Rectified Linear Unit	8
CRNN	Convolutional Recurrent Neural Network	8
TCN	Temporal Convolutional Network	8
GRU	Gated Recurrent Unit	8
Tanh	Hyperbolic Tangent	8
FFT	Fast Fourier Transform	11
DCT	Discrete Cosine Transform	11
RNN	Recurrent Neural Network	12
BVC	Biometrics Visions and Computing	15

ANN	Artificial Neural Network	18
MSE	Mean Squared Error	21
API	Application Program Interface	28
GUI	Graphical User Interface	28
ASP	Automatic Speech Recognition	29
CSV	Comma-Separated Value	29
US	United States	30
kHz	Kilohertz	35

Table of contents

General Introduction

The Artificial Intelligence (AI) field of study is a rapidly expanding discipline that permeates various aspects of our daily lives. From self-driving cars to digital assistants, AI plays a part in revolutionizing many domains of technology. One of these domains is voice processing, where advanced algorithms can extract meaningful insights from audio signals. The voice is a typical human feature that can reveal a wealth of information about an individual, such as their emotions, identity, and even their age. This inherent potential has naturally raised considerable interest in the latest surge of AI-driven applications that leverage voice analysis.

Age estimation has become a particularly important application in recent years, especially with the advancements in AI technology. When done accurately, it can significantly enhance personalized interactive voice response services and improve security through identity verification. Additionally, accurate age estimation can refine content restriction methods and optimize recommendation algorithms for a better user experience. It can have a significant impact on the fields of healthcare and social sciences as understanding vocal changes in relation to age can contribute to better diagnostic tools and demographic studies.

The human voice is a complicated and unique tool for communication that conveys several characteristics of an individual. An AI-based solution can accurately determine a person's age by analyzing voice features such as tone, pitch, frequency, and speech patterns. As deep learning (DL) models and techniques advance, their ability to process and interpret these vocal features becomes more effective. This process involves training a neural network (NN) on large datasets to recognize patterns associated with different age groups.

The primary objective of this study is to explore and develop an AI-based model that can estimate age from voice. This involves delving into the fundamental concepts and principles of AI and DL, searching for adequate data that can represent different demographics

and age groups impartially, and building, testing, and refining the voice-based age estimation system to achieve the highest quality.

This master project provides a detailed description of our journey from conception to the implementation of a voice-based age estimation AI system. The development of such a system leverages advanced DL networks to capture and analyze intricate vocal features that traditional methods might miss, demonstrating why this approach is a promising and innovative solution. By emphasizing the wide variety of effects and real-world applications of voice-based age estimation, in addition to its technical components, the thesis advocates for the importance and potential of this technology.

The report is organized into three main chapters. Chapter 1, titled 'General Context', presents the background and context of the study, including the specifics of voice features and the significance and applications of voice-based age estimation. Chapter 2, titled "Deep Neural Networks", is dedicated to general information on deep neural networks (DNNs) and their role in AI and DL. Chapter 3, titled "Implementation", details the practical steps taken to construct and develop the AI system, including data collection, model training, and evaluation. We then conclude this report with a general conclusion that summarizes the main achievements of our work.

Chapter 1

General Context

1.1 Introduction

Speech stands as one of humanity's most important and complex activities, playing a significant role in daily interactions. It is often used as a basis for determining non-linguistic information about the speaker, including their age or emotions. In recent years, AI systems have increasingly started to leverage voice segments to determine a speaker's age, enabling a range of applications from age-based content restrictions to personalized product and service recommendations.

In this chapter, we delve into the foundational concepts behind our project, gaining a comprehensive understanding of the qualities of the human voice, the methodology of age estimation, and the structure of a voice-based age estimation system.

1.2 Uniqueness and development of the human voice

The human voice has always served as a fundamental tool for communication, while also playing a significant role in identifying a person’s age and gender. In recent years, voice-based biometrics analysis has emerged as one of the most significant technologies in AI. The voice is often chosen as a measure of identification and authentication because of the simplicity with which it can be captured and processed. It is also considered to be a less intrusive method of verification compared to its facial recognition counterpart.

The uniqueness of the individual’s voice is what makes research on this topic possible. At its core, the human voice emerges as air passes through the vocal folds from the lungs to the mouth and nose [1]. Key characteristics such as tone, timber, pitch, range, articulation, and

frequency are developed from this process, and are in turn used to differentiate the speaker's age and gender.

1.3 Age estimation

In real-life situations, age estimation is done for multiple reasons. For instance, in the medical field, it can be used in tailoring specific treatments to different age groups, while in social interactions, it ensures appropriate behavior when addressing elders or peers. Otherwise, age estimation can assist law enforcement with missing persons cases or in criminal investigations where limited information is available. It is also used in a wide range of other industries, such as market research and targeted advertising.

With the rise and rapid development of AI, the field of age estimation systems was bound to be discovered. To strengthen the security and privacy measures of the individual, it was first attempted by analyzing the growth of facial attributes in face recognition systems. Facial age estimation is the process of training a model to return a value representing the age of a person, it can either be an age range (classification problem) or an exact number (regression problem) [2]. NNs have also been utilized in facial age estimation tasks and have so far returned the best classification results [3].

Voice-based age estimation represents another significant field of research. Systems are trained to analyze speech signals and extract relevant features that can be associated with aging. In voice-based age categorization tasks, Machine Learning (ML) and DL models have achieved accuracies around 80% or higher in previous studies [4] [5]. These advancements have wide-ranging effects, particularly in enhancing security procedures and improving Interactive Voice Response (IVR) systems. They also refine recommendation algorithms, facilitating age-based targeted advertising, personalized content delivery for specific age groups, and age-specific customer segmentation for marketing purposes.

1.4 Previous work

The evolution of voice-based age estimation research has unfolded over several years. The research done on the topic often utilizes DL models, as they have demonstrated discriminative capabilities in analyzing biometric data. Earlier studies relied on ML models like the Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) classifier. For instance, in 2007, the GMM classifier was applied to a Mel-Frequency Cepstral Coefficient (MFCC) feature vector and returned 94.6% in accuracy [4].

Additionally, in a 2011 study, an SVM classifier was used with both MFCC and PLP (Perceptual Linear Predictive) feature vectors and yielded 5.89% and 8.84% in error rates respectively [6].

In DL, different DNN embedding architectures have been assessed for voice-based age estimation, mainly the x-vector and d-vector embedders within Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) frameworks in a 2021 study [7]. The most accurate results were obtained using the CNN Quartznet-based x-vector model, which was pre-trained and tested on three different diverse datasets, achieving a Mean Absolute Error (MAE) of 5.12 for males and 5.29 for females.

In 2022, several network architectures were evaluated on one large dataset. These included a CNN with a 2D Convolutional layer and a Rectified Linear Unit (ReLU) activation function, a Convolutional Recurrent Neural Network (CRNN) that incorporate recurrent layers for finding patterns in sequences, and a Temporal Convolutional Network (TCN) that can model sequences using convolutional layers. Additionally, a Gated Recurrent Unit (GRU) model with Hyperbolic Tangent (tanh) and sigmoid recurrent activation functions was also examined. The results indicated that DNNs with structures that allow finding patterns in temporal sequences are better suited for the speech-based age estimation task, with the TCN and CRNN networks outperforming the CNN one [8].

Also in 2022, the company Privately SA, in partnership with VerifyMyAge, launched the first voice-based age verification application: VoiceAssure [9]. Users would read a randomly generated sentence correctly from their screen for the program to estimate their age and effectively manage their access to appropriate content.

1.5 Problem definition

In this project, we aim to develop a voice-based age estimation system that can accurately determine a speaker's age range from a brief speech sample. Our primary motivation is to tackle the safety concerns arising from unrestricted internet access, and to enhance user's experience by personalizing online interactions and content based on age demographics.

The project workload includes collecting a diverse dataset of speech segments labeled with the speaker's age, developing, and training a DNN model to extract important features associated with aging from the data, and evaluating the model's performance across various demographic groups to discern potential variations.

1.6 Voice features

The accuracy of age estimation heavily relies on the selection of features extracted from voice signals. The impact of various life stages on vocal characteristics can be described as follows:

- Childhood: Children's voices tend to have a higher pitch than adults due to their smaller larynx (voice box) and thinner vocal cords [10].
- Puberty: Hormonal changes during this phase lead to the growth of the larynx as it moves lower down the neck, along with the lengthening and thickening of vocal cords (Figure 1.1). These changes typically occur between the ages of 12 and 16 for boys and between 10 and 14 for girls.

- Early Adulthood: At this stage, the larynx reaches its mature size, and the vocal cords stabilize. Adult men typically have deeper voices than adult females due to their longer vocal cords.

- Aging: As individuals age, subtle changes may occur that weaken the voice. The vocal cords may dry out and lose muscle tone, flexibility, and elasticity. Additionally, the muscles of the larynx can decline, becoming thinner and weaker, and the laryngeal cartilage may calcify [10].

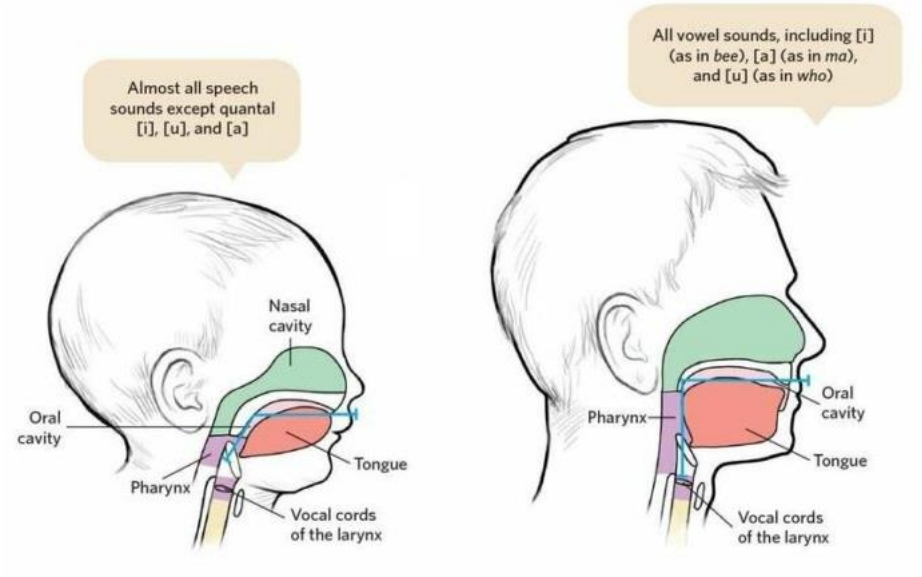

Figure 1.1 - The Larynx of a child and an adult [11].

1.7 MFCC feature extraction algorithm

MFCCs are a set of features extracted from the mel-spectrogram of an audio signal. They capture relevant information about the spectral content of the signal, such as frequency information and speech articulation changes, emphasizing important features while reducing dimensionality [6]. The MFCC feature extraction process is as follows:

- The speech signal is first divided into short frames, typically spanning 20–40 milliseconds each, where higher frequencies are emphasized.

- A Hamming window function is applied to each frame to reduce spectral leakage.
- The time domain frames are then transformed into frequency frames using the Fast Fourier Transform (FFT) algorithm.
- These frames are then processed by the Mel filter bank (which approximates human perception of pitch) that then outputs logs energies of the spectral features.
- These log energies are then processed by the Discrete Cosine Transform (DCT) that turns them into a set of cepstral coefficients, and the MFCC vector is obtained [12] (Figure 1.2).

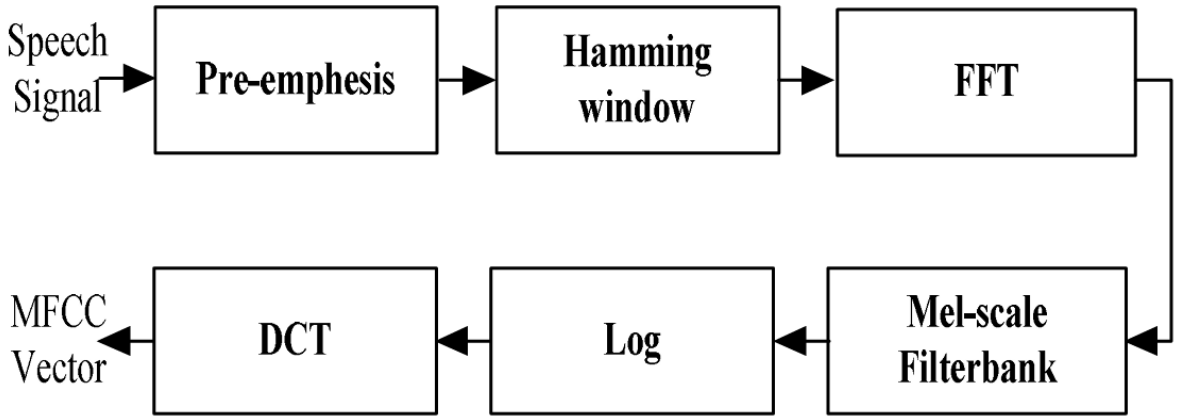

Figure 1.2 - The MFCC process.

1.8 Voice-based age estimation model architecture

The architecture of a DNN system can be split into two phases: training and testing. The model gains the ability to identify patterns in the input data during the training phase. To do this, it modifies its internal parameters to reduce the discrepancy between its predictions and the actual output labels, using a loss function to measure the accuracy of the model's predictions. In the testing phase, the trained model's performance is evaluated on unseen data, which is essential to comprehending the model's ability to generalize.

1.8.1 Input data

Outside of the actual system, there are several essential parameters to consider when collecting data. These include diversity in the age and gender of the speakers, as well as their languages, accents, and dialects. Additionally, the quality of the audio samples and the content of the speech are also important features.

The raw audio speech signals are preprocessed before being fed into the DNN model; this includes getting rid of any background noise, normalizing the audio levels, and potentially applying techniques like noise reduction or filtering to enhance the signal quality.

1.8.2 Feature extraction

The feature extraction process is designed to mimic the human auditory system. Pre-defined functions are used to convert the audio signal into a set of acoustic observations and feature vectors, capturing relevant information vital for estimating age. Its main aim is to minimize the size of valuable data for the classifier.

During this process, the input signal is empathized, filtered, and divided into smaller frames. From each frame, a set of features are extracted, often including MFCCs that capture the phonetic content of the signal, as well as other spectral characteristics. The features are then normalized to have a consistent scale, and then a sequence of feature vectors is passed on to the classifier for the age estimation.

1.8.3 Classification

To identify the age range of the speaker a classification algorithm is employed. DL algorithms, such as CNNs, RNNs (recurrent neural networks), and LSTM units, have been particularly effective in this context. The biggest challenge is selecting the most suitable classification algorithm that will achieve the highest accuracy while taking into consideration the quality of the extracted features, as well as the size and representativeness of the training dataset.

The classification stage involves training a DNN model on labeled (supervised) data to identify patterns in the feature vectors that indicate the age range of the speaker. Additionally, a validation set is incorporated to fine-tune the model and identify overfitting and underfitting problems early in the process. Subsequently, the trained model is employed to estimate the age range of unseen speakers during the testing phase. The performance of the classifier is entirely dependent on the effectiveness of the previous stages.

1.8.4 Evaluation

To assess the classifier's performance, several metrics can be used:

- Recall: This metric measures how often a model correctly identifies positive instances (true positives) in relation to all actual positive instances in a dataset. It is calculated using the equation (1):

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (1)$$

- Accuracy: This metric calculates the proportion of correct predictions over the total number of predictions. It is calculated using the equation (2):

$$\text{Accuracy} = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives} \quad (2)$$

- Precision: This metric indicates whether a positive prediction made by the model is accurate or not. It is calculated using the equation (3):

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

- F1 Score: This metric calculates the reliability of the model by finding the harmonic mean of precision and recall. It is calculated using the equation (4):

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

1.9 Datasets

Numerous datasets comprising speech audio samples labeled with speaker age are accessible for research purposes. We evaluated their features to ascertain their relevance for the objectives of this project.

1.9.1 Mozilla Common Voice

The Common Voice dataset is a multilingual collection of voice recordings contributed by volunteers, typically reading out sentences from public domain sources. It is highly diverse, encompassing a wide range of languages, accents, and ages, recorded in various conditions. The dataset consists of multiple versions collected over the years, with contributions in 124 different languages amounting to over 20,000 of validated hours. Each of these versions focus on various languages and regions, providing metadata about the age, gender, and accent of the speaker, along with transcriptions of the audio [13].

Several versions of it have been used to train and test voice-based age estimation systems in multiple studies, with models achieving state-of-the-art performance when trained on the large and diverse Common Voice dataset and then fine-tuned and tested on smaller and more local datasets [7]. Though training and testing a DL model on a locally developed Korean speech dataset outperformed the same model trained and tested on the Common Voice dataset with a 24% margin, this difference can be attributed to the biases that can develop in a model when using data from one specific demographic, that may not be representative of the broader population [5].

1.9.2 VoxCeleb

VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. It has two releases: VoxCeleb1, which consists of more than 100,000 utterances from 1251 celebrities, and VoxCeleb2 which contains over 1 million utterances from 6,112 celebrities.

This dataset is mostly gender balanced (males comprise 61% and females 39%), and it spans over a wide range of ethnicities, nationalities, and ages. It includes metadata with demographic information such as age, gender, and accent. The length of its voice segments is small, spanning just between 4 and 20 seconds.

Its first release, VoxCeleb1, has been used in a former study for pre-training the model's embedder alone on speaker's identification, and the model returned its best results for age estimation afterwards [7]. It also scored practically equally and effectively on gender dependent and independent systems that estimate age from short speech utterances without depending on the text [14].

1.9.3 Biometrics Visions and Computing (BVC)

The Biometrics Visions and Computing (BVC) voice dataset consists of utterances from 526 individuals, 336 of which are males and 190 are females. The number of utterances is 3,964, consisting of 2,149 male and 1,815 female voice utterances.

One to five voice recordings in English language and native languages were acquired from each of the subjects. Twenty-eight different native languages make up the native language set. The voices in this dataset are challenged with some background noise that does not seriously affect the audios.

The results of using this dataset for gender classification showed that the mother tongue or first language, intonation variations, language variety in the training and testing set influence the model's results [15].

1.10 Benefits and challenges of voice-based age estimation

The benefits of a voice-based age estimation system are various. This kind of application can be used for identification purposes, and to enhance security while considering user's privacy and requiring limited and non-invasive interaction. Also, because of its real-time analysis aspect, it can provide immediate feedback, which would be utilized in customer

service where understanding the age range of the caller is essential to suit their needs. Furthermore, it can be used in content customization, where the system can recommend or restrict content, and finally in recommendation algorithms to enhance user's experience.

The challenge with a speech-based age estimation system is that a significant percentage of its errors can be attributed to environmental and emotional factors. Conditions such as smoking, and throat cancer can complicate the task of accurately estimating the age from voice alone. Additionally, there's a lack of large and diverse datasets in the field, with most databases only focusing on one language or region of the world, and almost all of them having some disbalance in gender distribution. Another setback that needs to be mentioned is that adjacent age ranges can have very similar voice features, making it harder to differentiate between them.

1.11 Conclusion

In this chapter, we have presented a comprehensive overview of voice-based age estimation. Starting with understanding the human voice functioning and development, and the voice features that play a role in age estimation. We also showcased the general architecture that our model would follow, the metrics that it can be evaluated on, and the different datasets that have been used in similar studies.

In the next chapter, we will delve more into the specifics of a voice-based age estimation model and provide more information on the different DL methods and architectures that could manage to estimate the age of a speaker.

Chapter 2

Deep Neural Networks

2.1 Introduction

In recent years, the fields of AI, ML and DL, have swept in and revolutionized many domains, ranging from image and video processing to strengthening cybersecurity, healthcare, agriculture and more. Age estimation stands out as a specialty that has seen significant advancements due to these technologies.

This chapter introduces fundamental concepts in DL, including DNNs, CNNs, RNNs, and their various types of networks, and defines their roles and functionalities in the context of age estimation from speech.

2.2 Artificial Intelligence (AI)

The AI label can be assigned to any computer system or machine that is capable of imitating human intelligence in tasks such as problem-solving, recognizing sounds or images, or decision-making. AI systems learn how to do so by processing massive amounts of data and learning to identify features to make decisions [16].

The field of AI powers a diverse range of applications, ranging from medical diagnosis systems enhancing the healthcare industry to autonomous vehicles navigating city streets. ML and DL, integral components of AI (Figure 2.1), enable these systems to analyze vast amounts of data, identify patterns, and make informed decisions and predictions.

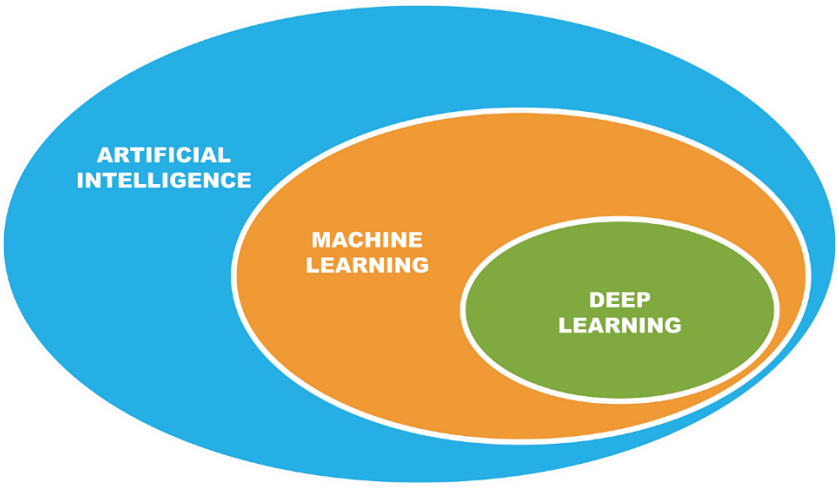

Figure 2.1 – The relationship between AI, ML, and DL.

2.3 Machine Learning (ML)

ML is a subset of AI that's whole purpose is to create algorithms that can learn and improve without human intervention. These algorithms identify patterns in datasets and use these patterns to improve their performance over time without being explicitly programmed to do so. The data that an ML model can be trained includes anything from words, numbers, images, statistics, sounds, and more [17].

2.4 Deep Learning (DL)

DL is a branch of ML that leverages multi-layered DNNs to process complex data and make predictions. DL technology originated from Artificial Neural Networks (ANNs) and is widely used in various fields, including healthcare, cybersecurity, computer vision, audio analysis, natural language processing, and more.

DNNs are widely used in voice-based age estimation research due to the inherently high-dimensional and complex nature of speech signal data. They excel at automatically learning intricate patterns and representations from such data using advanced techniques and model architectures.

2.4.1 Artificial Neural Network (ANN)

An ANN is composed of a hierarchical arrangement of interconnected neurons, comprising an input layer, one to two hidden layers, and an output layer. Typically, ANNs are employed for simpler computational tasks like handwriting recognition and basic image classification. In contrast, DNNs are utilized for more complex tasks such as speech recognition, computer vision, and natural language processing. DNNs have a more advanced hidden layer architecture, with layers such as convolutional, recurrent, pooling, and dense layers that are specifically designed to address the complexity of the problem at hand [18] (Figure 2.2).

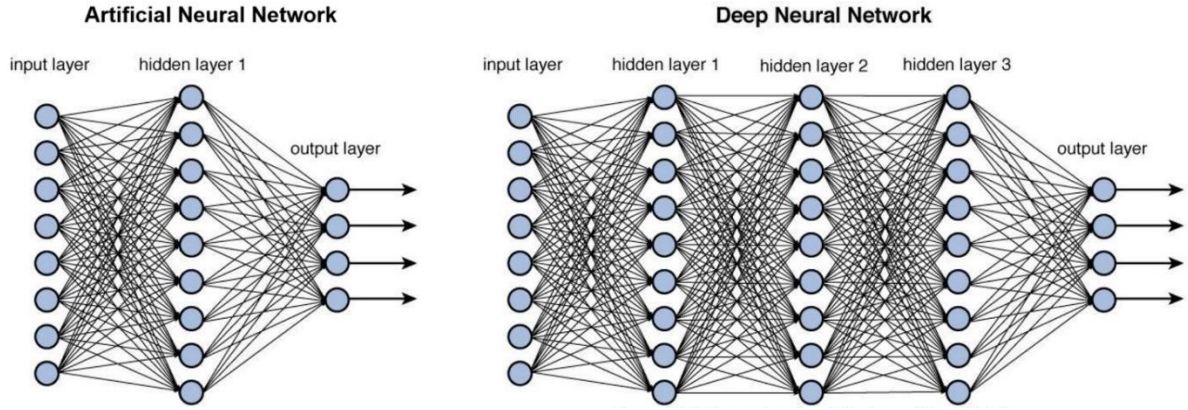

Figure 2.2 - ANN vs DNN [18].

In this layered structure, the perceptron, an artificial neuron inspired by the function of a biological neuron and initially introduced as a supervised learning algorithm for binary classification, operates by taking a set of input features x_i from n dimensions, calculating a weighted sum w_i, adding a bias to them b, and producing an output y based on a non-linear activation function f() [19] as seen in (5).

$$y = f(\sum_{i=1}^{n} \quad w_i x_i + b) \tag{5}$$

In practice, single perceptron models are limited to classification problems with linearly separable classes (Figure 2.3). As a result, they are frequently utilized as building blocks for ANNs and DNNs to handle a wider range of problems. A DNN essentially consists of multiple

layers of interconnected neurons, each of which generates a series of real-valued activations that's passed forward for the target outcome.

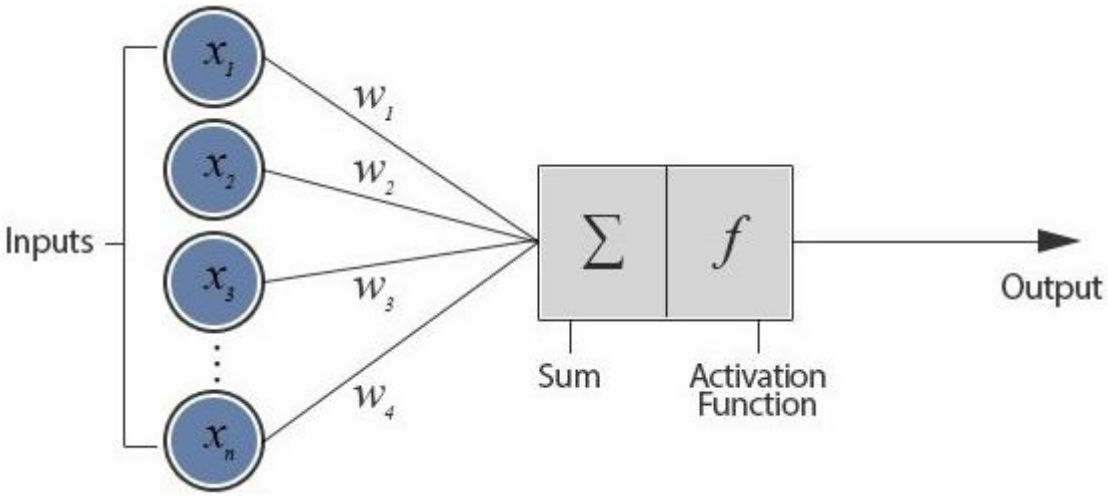

Figure 2.3 - Single layer perceptron example.

2.4.1.1 Activation function

An activation function shapes the output of each neuron by introducing non-linearity, enabling the network to learn and extract complex patterns and make accurate predictions. There are several activation functions available; the most common include ReLU, Tanh, Sigmoid, and Leaky ReLU, where 'z' represents the input (Figure 2.4) [20].

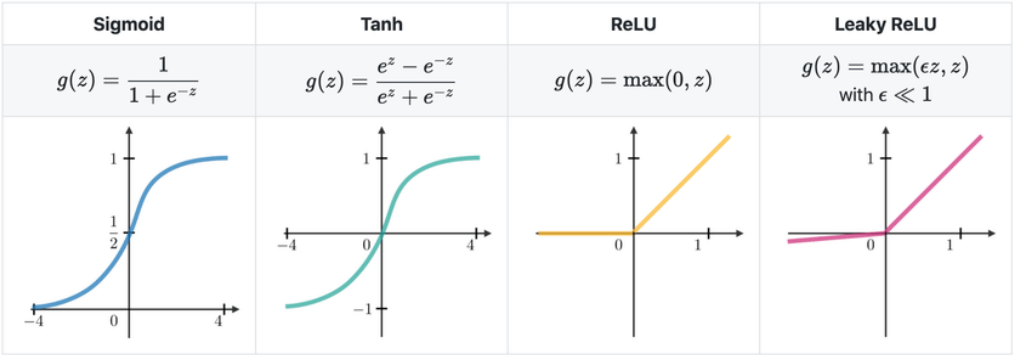

Figure 2.4 - Activation functions formulas and graphs [20].

2.4.1.2 Weights and bias

In NNs, weights represent the strength of the connections between neurons, while bias is used to shift the activation function's input-output curve. During the learning process, the network engages in iterative forward propagation (making predictions) and backward propagation (updating weights and biases based on prediction errors) with the goal of minimizing loss and enhancing accuracy [21].

2.4.1.3 Loss function

Loss functions are used to measure the discrepancy between the predicted values and their real counterpart, thereby revealing how well a network is performing and guiding the optimization process [22]. Examples of such functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss (Log Loss) for classification.

2.4.1.4 Regularization

A recurring problem in NN models is overfitting, where the network performs exceptionally better with the training set than returns a lower accuracy with the validation and testing sets. Regularization is a method that’s often used to dissipate this issue. By adding a penalty term to the loss function, it discourages the model from assigning too much importance to certain features, enabling it to generalize to new data better [23].

2.4.2 Convolutional Neural Networks

CNNs are a class of deep feed-forward networks that take input images and extract relevant features to efficiently identify and classify images. Its two main components are convolution and pooling layers (Figure 2.5). The convolution operation uses multiple learnable filters (kernels), which are small windows that slide over input data to extract features (feature map) that highlight specific patterns from datasets, through which their corresponding spatial information can be preserved, and output the transformed input into the next layer.

The pooling operation is used to reduce the dimensionality of feature maps from the convolution operation. It does that by combining the outputs of neuron clusters at the previous

layer into a single neuron in the next layer, performing down-sampling and dimensionality reduction to address Overfitting issues. In CNN, max pooling and average pooling are the most common pooling operations used, and ReLU is the common choice for the activation function to transfer gradients in training by backpropagation [24].

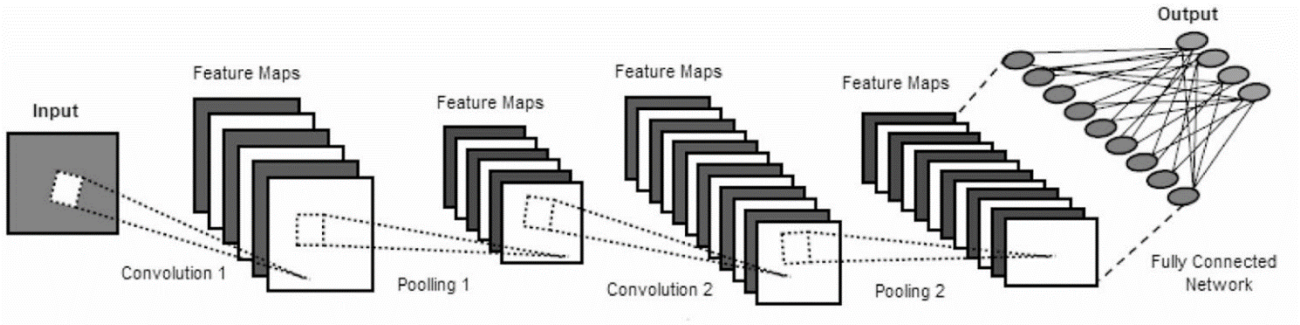

Figure 2.5 - An example of a CNN structure.

2.4.3 Recurrent Neural Networks

RNNs are a class of DNNs that can process time-series data and other sequential data by calculating their internal memory using the equation in (6):

$$y_t = f(W \cdot x_t + V \cdot y_{t-1} + b) \quad (6)$$

Here, t signifies the current time step in sequence, x_t represents the input at time t, y_{t-1} is the output from the previous time step, W and V the weight matrices for x_t and $y_{t-1,}$ b is the bias, f denotes the non-linear activation function, and y_t the output at time t.

RNNs can apply this same operation to every element in a series by employing a loop to capture temporal dependencies in the data (Figure 2.6).

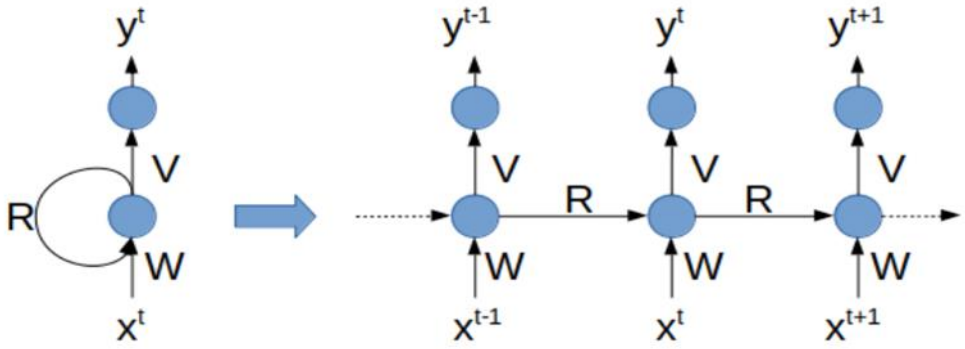

Figure 2.6 - RNN schema: recurrent version, and unfolded version.

Traditional RNNs are limited in their ability to manage long sequences due to their short-term memory, which allows for gradients to vanish as they are backpropagated from the output layers to earlier layers [24].

2.4.4 Long-Short Term Memory

LSTM unit is an advanced variant of RNNs that memorizes long-term dependencies of time-series data. It consists of a series of gates and cells that cooperate to control the information flow (Figure 2.7), which includes an input gate that update the internal state of a cell based on the current input and the previous internal state, a forget gate that decide what to forget from the previous internal state, and an output gate that regulates the influence of the cell state on the system.

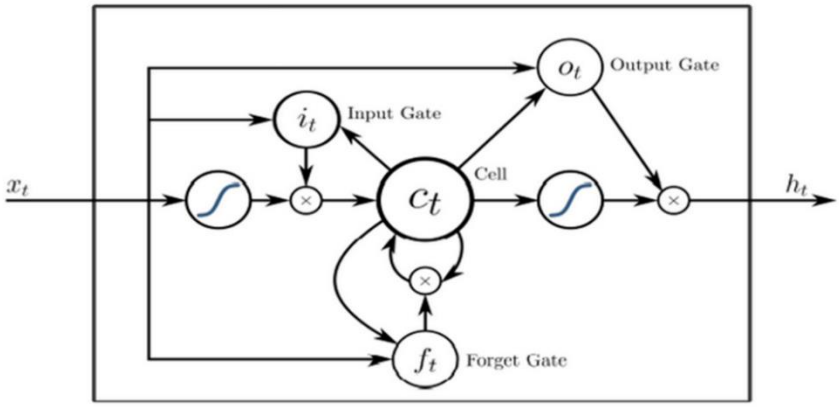

Figure 2.7 - LSTM Unit Base.

LSTM units are effective at managing information flow over extended time periods better than the average RNN [25]. This skill is especially useful for voice-based age estimation tasks, in which the model needs to identify minute temporal patterns in speech that correspond to vocal changes associated with aging. Because of its excellent generalization capabilities and ability to handle sequential data, LSTM is a good fit for tackling the difficult issues involved in age prediction from voice.

2.5 Why use DNNs?

DNNs offer a multitude of compelling advantages that make them highly beneficial to work with. These advantages include:

- Hierarchical Representation: DNNs can automatically learn hierarchical representations of data and are able to recognize relevant features directly from the raw data.
- Scalability: these networks are adept in processing large datasets. As the network depth expands, they can capture more intricate and complex patterns that shallower networks might miss.
- Non-Linearity: DNNs are flexible in their ability to model non-linear relationships between inputs and outputs, which is crucial when dealing with complex data.
- End-to-end Learning: A DNN model, during the training process, learns to extract features and make predictions simultaneously which leads to better performance from the jointly optimized system.

2.6 Objectives and constraints

The objective of this study is to construct a rapid and precise system that can predict age from the voice. This process involves deepening our understanding of various DNN methods and architectures to enhance effectiveness, while exploring innovative voice analysis and processing methods to achieve a robust model.

The constraints of this work include the computational complexity involved in processing voice features, such as extracting and transforming audio data into suitable formats for DNN models. Additionally, acquiring substantial and consistent voice data is challenging, especially when taking into consideration the disparity in all the age groups and demographic present. Furthermore, there are inherent difficulties related to bias, overfitting, and underfitting problems in model training, which are critical considerations in the model assessment phase.

2.7 Conclusion

In conclusion, this chapter provides an overview of key concepts in DL, specifically DNNs, and the different architectures that they can fold into and their distinct approaches to fulfil the task of age estimation from speech.

Moving forward, further research and exploration into the specific structure of an age predicting model, along with the data treatment that precedes it, will be explored in the next chapter.

Chapter 3

Implementation

3.1 Introduction

In this chapter, we delve into the practical aspects of implementing our solution. We begin by taking a complete inventory of the physical and intangible materials at our disposal, which allows us to get a sense of the computational capacity available to us.

Next, we break down the steps involved in our data treatment process. Detailing how we transformed raw audio speech segments into a more accessible format that can be understood by the model. We then reveal the architecture of our prediction model, train it to the best of its ability, and evaluate its performance using new, unseen data.

3.2 Development environment

3.2.1 Hardware

In this study, we have used a computer equipped with:

- Processor (CPU) type: Intel Core i7-10510 with a base clock speed of 1.80GHz and can turbo boost up to 2.30GHz.
- RAM: 16.0 GB.
- External hard drive: 1000 GB of storage capacity.

3.2.2 Software

- Operating System: Windows 11 Professional (64-bit)
- Source Code Editor: Visual Studio Code (Version 1.89.0, User setup)

- Python: Version 3.11.4

3.3 Development tools

3.3.1 Python

Python is an open-source all-purpose high-level programming language with several applications, including web development, data science, software development, object-oriented programming, and more [26].

Python has a comprehensive standard library that's made available across all its implementations. It's dynamically typed, garbage-collected, and includes both structured and functional programming [27].

3.3.2 Jupyter Notebook

An open-source and interactive computing environment with a wide selection of programming languages, including Python, R, and Julia. It includes JupyterLab, a web-based interactive development environment, that allows developers to manage coding projects in data science, scientific computing, and ML.

The classic notebook interface is not much different. As the original web application for developing and sharing computational documents it offers a simple, efficient, and document-centric experience with the ability to write code and include visualizations easily [28].

3.3.3 Scikit-learn

An open-source ML library designed to integrate with numerical and scientific Python libraries. It encompasses several classification, regression, and clustering algorithms. It also supports tasks such as dimensionality reduction, feature selection algorithms, grid search, and cross validation [29]. Scikit-learn also provides a selection of predefined functions for model evaluation metrics, catering to both regression and classification tasks.

3.3.4 TensorFlow and Keras

TensorFlow is an open-source, end-to-end AI platform and library that provides a flexible framework that can be used to develop and train complex model architectures with the Keras API (Application Program Interface). It supports both functional and sequential model building approaches and provides tools for data processing and preprocessing. Additionally, TensorFlow offers model deployment services, along with immediate model iteration and faster debugging through eager execution [30].

Keras is an open-source DL library for Python, that primarily serves as an interface for the ML library TensorFlow. It offers simplicity, flexibility, powerful performance, and scalability. With its high-level APIs, it allows developers to define NNs architecture with ease. Additionally, Keras provides various pre-defined layers such as dense (fully connected), convolutional, recurrent, pooling, and more, with the option for customization. It also handles model compilation, training, testing, transfer learning using pre-trained models, and data augmentation [31].

3.3.5 Librosa

A python package for audio processing, it can be used to extract relevant features like MFCCs and spectral characteristics from voice segments. It supports various formats, including MP3, WAV, FLAC, and more [32].

3.3.6 Matplotlib

An object-oriented plotting library for Python. It is used to create static, animated, and interactive visualizations with the option to customize the styles and layouts, as well as providing an API for embedding plots into applications using Graphical User Interface (GUI) toolkits (e.g. confusion matrix).

3.4 Data preprocessing

3.4.1 Data collection

The dataset we used for this project is the Mozilla Common Voice 2018 Corpus, released on Kaggle by the Mozilla Org account[1]. This version has been utilized in several voice-based age estimation projects on the website, where the prediction models SVM [33] and the Random Forest Classifier [34] were utilized and returned an average F1 score of 0.689 and 0.614 respectively.

The dataset includes 380,000 entries, which corresponds to 500 hours of recordings stored in MP3 format. Its primary purpose is to create Automatic Speech Recognition (ASR) systems, which is why it has been divided into three subsets: 'valid', 'invalid', and 'other'.

In the 'valid' subset, each audio clip has been reviewed by at least two individuals, with the majority confirming that the transcribed text matches the audio content. While in the 'invalid' subset, most reviewers decided that the transcription does not match the audio. The 'other' subset comprises audio clips that either have less than two reviews or have an equal number of upvotes and downvotes from reviewers.

The 'valid' and 'other' subsets are further split into training set, testing set, and validation set. Additionally, the dataset folder includes CSV (Comma-Separated Value) files for each subset, providing the attribute information for each entry. The purpose of each column is outlined in Table 3.1.

[1] Installed from the Kaggle Website: https://www.kaggle.com/datasets/mozillaorg/common-voice.

Table 3.1 – The Common Voice dataset columns.

Column Name	Description	Data Type	Categories
Filename	Relative path of the audio file.	String	
Text	Supposed transcription of the audio.	String	
Up_votes	Number of people who said the audio matches the text.	Integer	
Down_votes	Number of people who said the audio does not match the text.	Integer	
Age	Participant age.	Categorical	Teens (< 19), twenties (19-29), thirties (30 -39), fourties (40-49), fifties (50-59), sixties (60-69), seventies (70-79), eighties (80-89), nineties (> 89).
Gender	Participant gender.	Categorical	Male, female, other.
Accent	Participant country of origin.	Categorical	US (United States), England, Australia, Indian, Canada, Malaysia, Ireland, Bermuda, Scotland, African, New Zealand, Wales, Philippines, Singapore, Hongkong, South Atlantic.
Duration	Full length of the participant speech segment.	Floating-point number	

The dataset CSV files were imported for model training using the subsequent Python script in Figure 3.1.

```python
file_paths = [
    "D:/AMINA/PFE24/datasets/commonvoice/cv-valid-train.csv",
    "D:/AMINA/PFE24/datasets/commonvoice/cv-other-train.csv",
    "D:/AMINA/PFE24/datasets/commonvoice/cv-invalid.csv",
    "D:/AMINA/PFE24/datasets/commonvoice/cv-valid-dev.csv",
    "D:/AMINA/PFE24/datasets/commonvoice/cv-other-dev.csv",
    "D:/AMINA/PFE24/datasets/commonvoice/cv-valid-test.csv",
    "D:/AMINA/PFE24/datasets/commonvoice/cv-other-test.csv",
]
dfs = []
for i in file_paths:
    df = pd.read_csv(i)
    dfs.append(df)

df = pd.concat(dfs, ignore_index=True)

output_file_path = os.path.join("D:/AMINA/PFE24/application", "data.csv")
df.to_csv(output_file_path, index=False)
```

Figure 3.1 - Python script to import all csv files from the Common Voice dataset.

3.4.2 Data cleaning

Data preprocessing plays an important role in DL, aiming to improve model performance. It includes techniques such as comprehending, cleaning, and transforming data to make it more suitable for analysis, along with feature creation and selection to reduce misleading results.

In this section, we examine the different demographics that comprise the Common Voice dataset to gain a better understanding of the population under study and make slight modifications to better prepare the data for model training.

Given that the Common Voice dataset was not originally intended for age estimation, a substantial portion of speech segments lacking age target variables had to be discarded. The distribution of the remaining rows among the eight age classes is unbalanced, with a notable prevalence of speakers in their twenties and thirties, while there are considerably fewer speakers in their seventies and eighties (Figure 3.2). This imbalance in age distribution may potentially impact the model's performance in predicting individual age classes.

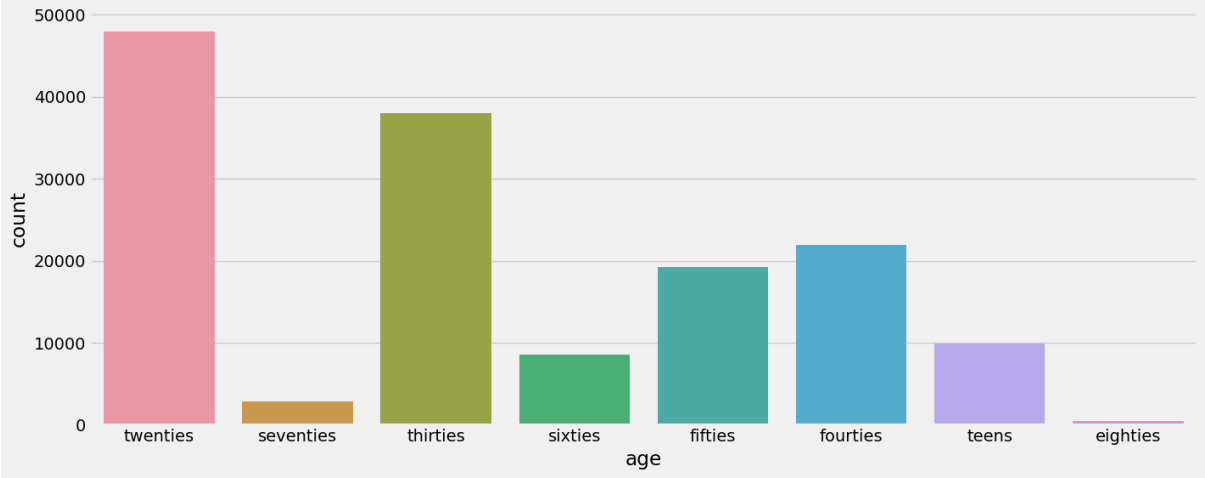

Figure 3.2 - The age distribution in the Common Voice dataset.

Additionally, the columns related to speech transcription systems were removed as part of the data cleaning process (Figure 3.3). This decision was made because, although a speaker's age can sometimes be estimated from their vocabulary, in the Common Voice dataset, all speakers are reading assigned text from public domain sources. This deduction also allows us to utilize data from all three folders—valid, invalid, and other—thereby expanding our dataset for training the model.

```
    df = df[["filename","age","gender","accent"]]
    df.info()
[119]  ✓ 0.0s
...  <class 'pandas.core.frame.DataFrame'>
     Index: 149021 entries, 5 to 380366
     Data columns (total 4 columns):
      #   Column    Non-Null Count   Dtype
     ---  ------    --------------   -----
      0   filename  149021 non-null  object
      1   age       149021 non-null  object
      2   gender    148487 non-null  object
      3   accent    126628 non-null  object
     dtypes: object(4)
     memory usage: 5.7+ MB
```

Figure 3.3 - Python script demonstrating the removal of irrelevant columns from dataset.

After removing missing values from the remaining gender and accent columns, we were left with a dataset containing 126,108 usable rows. Most speakers in this dataset originate from the US, followed by England. Additionally, it was observed that male speakers predominated across nearly all accents (Figure 3.4).

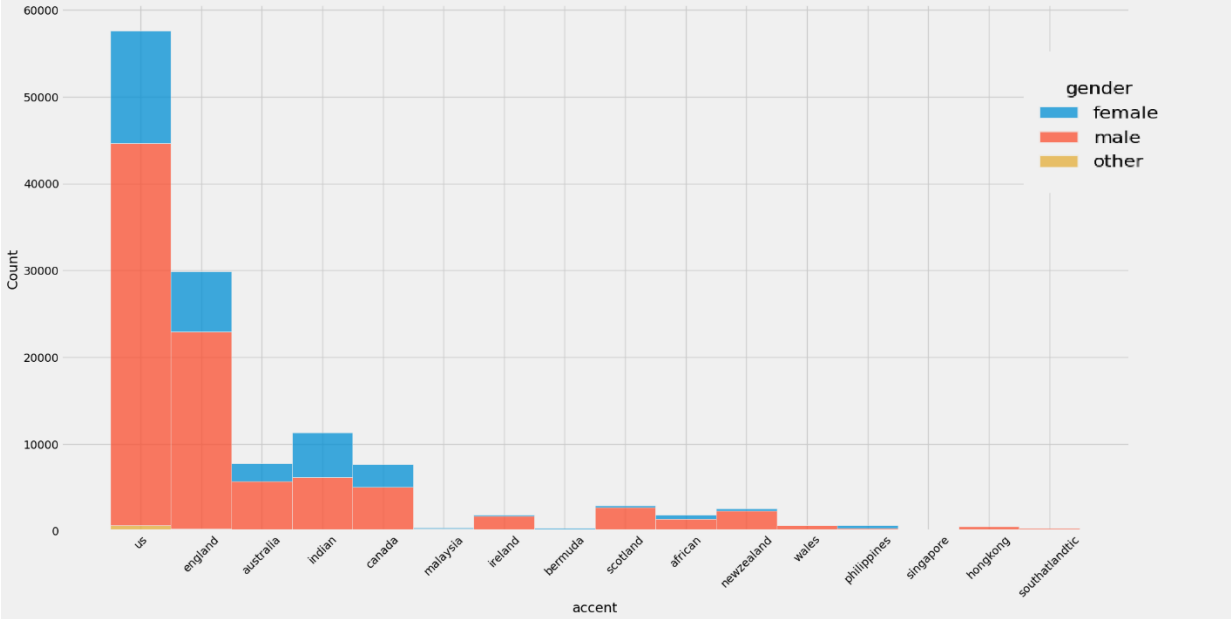

Figure 3.4 - The accent and gender distribution in the cleaned Common Voice dataset.

Furthermore, to prevent the model from assuming ordinal relationships between the categories of our main columns—age, gender, and accent—, and to ensure that each category is treated independently, we used One-hot encoding (Figure 3.5).

This technique represents categorical variables as binary vectors by creating a new binary column for each category. For example, in the gender column, “male” becomes [1, 0], and “female” becomes [0, 1]. However, this approach could lead to increased dimensionality and sparsity, making the model more complex and slower to train [35].

```
categorical_cols = ['gender', 'accent']
numerical_cols = [col for col in df_features.columns if col not in ['age', 'gender', 'accent']]

# One-hot encoding the categorical columns
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(), categorical_cols)
    ],
    remainder='passthrough'#numerical data passes unchanged
)

#Encoding the 'age' column using LabelEncoder() to convert the age categories into numerical labels.
X_preprocessed = preprocessor.fit_transform(df_features.drop(columns=['age']))
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(df_features['age'])
```

Figure 3.5 - Python script to one-hot the categorical columns: age, gender, and accent.

3.4.3 Feature extraction

In the next step of our process, we extracted key characteristics from the speech segments:

- 20 MFCCs: These coefficients provide information about the voice quality and vowel recognition through the vocal tract's shape and the resonant frequencies of the voice. Using 20 MFCCs works as a standard for information richness—while avoiding underfitting and overfitting—and computational efficiency, along with having former studies attain notable results with them [36] [6].

- Spectral Centroid: A measure used in digital signal processing to indicate where the center of mass of the spectrum is located. In speech segments it can provide the average frequency content, with higher values signifying that the voice has more energy at higher frequencies [37].

- Spectral Bandwidth: Indicates how wide or narrow the frequency distribution is in a sound signal. A person with a wider spectral bandwidth may have a voice that contains energy across a broader range of frequencies and sounds more resonant or rich [38].

- Spectral Rolloff: Represents the frequency below which a specified percentage of the total spectral energy. By calculating the rolloff frequency of each frame in a speech signal

we gain insight into the speaker's hearing sensitivity, which may be more pronounced in older individuals [39].

We used the Librosa library to extract these characteristics from each individual speech segment after sampling them down to 16kHz (kilohertz) to reduce computational load (as seen in Figure 3.6).

```
def feature_extraction(filename, sampling_rate=16000):
    path = "{}{}".format(ds_path, filename)  # Correct path construction
    features = list()
    audio, _ = librosa.load(path, sr=sampling_rate)

    gender = df[df['filename'] == filename].gender.values[0]
    accent = df[df['filename'] == filename].accent.values[0]
    spectral_centroid = np.mean(librosa.feature.spectral_centroid(y=audio, sr=sampling_rate))
    spectral_bandwidth = np.mean(librosa.feature.spectral_bandwidth(y=audio, sr=sampling_rate))
    spectral_rolloff = np.mean(librosa.feature.spectral_rolloff(y=audio, sr=sampling_rate))
    features.append(gender)
    features.append(accent)
    features.append(spectral_centroid)
    features.append(spectral_bandwidth)
    features.append(spectral_rolloff)
    mfcc = librosa.feature.mfcc(y=audio, sr=sampling_rate) #16GHz
    for el in mfcc:
        features.append(np.mean(el))

    return features
```

Figure 3.6 - Python script to extract voice features using the Librosa library.

3.4.4 Data split

To have an objective assessment of the model's performance, the dataset was split into three distinct subsets—training, validation, and testing—using the Scikit-learn Stratified k-fold function for unbalanced data (Figure 3.7). This technique allows us to preserve the class ration in all sets, so that the model does not overfit to a particular class distribution.

```
X_preprocessed = preprocessor.fit_transform(df_features.drop(columns=['age']))
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(df_features['age'])
X_train, X_temp, y_train, y_temp = train_test_split(X_preprocessed, y_encoded, test_size=0.2, stratify=y_encoded, random_state=42)
X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, stratify=y_temp, random_state=42)
```

Figure 3.7 - Python script to split unbalanced data with Stratified k-fold.

The training set includes 80% of the dataset and is used to train the model to recognize underlying patterns and features, the validation set includes 10% of the data and is used to evaluate the model after each training epoch and detect overfitting, the remaining 10% is the testing data that provides an unbiased estimate of how well the model will perform on new unseen data. The age distribution in each of the subsets is as depicted in Table 2.3.

Table 3.2 – Age Distribution and Dataset Split.

Age Groups	Age Range	Train	Validation	Testing
Teens	< 19	6431	804	804
Twenties	20 – 29	29305	3663	3663
Thirties	30 – 39	26156	3269	3270
Forties	40 – 49	15265	1909	1908
Fifties	50 – 59	14316	1789	1790
Sixties	60 – 69	6789	849	848
Seventies	70 – 79	2264	283	283
Eighties	> 80	360	45	45
Total of Instances		100,886	12,611	12,611

3.5 Model architecture

Considering our computational resources and previous research experimentation with different models [7] [14], we chose to deploy a prediction model utilizing RNN with LSTM cells for our classification problem. The detailed network architecture is outlined in Table 3.3.

Table 3.3 – Prediction model architecture summary.

Layer	Layer Size
LSTM_1	256
MaxPooling1D	–
LSTM_2	256
LSTM_3	128
Dropout	–
Flatten	–
Dense (ReLU Activation)	32
Batch Normalization	–
Dense (SoftMax Activation)	8 (number of classes)

We used the Adam optimizer to adjust the model's parameters during training, the sparse categorical crossentropy loss function to measure the difference between the predicted and real class labels, and the accuracy metric to evaluate the performance of the model. The training process was split into 10 epochs, each consisting of iterations with a batch size of 32 (Figure 3.8 and Figure 3.9).

```python
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_val = scaler.transform(X_val)
model = Sequential([
    LSTM(units=256, return_sequences=True, input_shape=(X_train.shape[1], 1)),
    MaxPooling1D(pool_size=2),
    LSTM(units=256, return_sequences=True),
    LSTM(units=128),
    Dropout(0.3),
    Flatten(),
    Dense(units=32, activation='relu'),
    BatchNormalization(),
    Dense(units=len(label_encoder.classes_), activation='softmax')
])

model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
history = model.fit(X_train.reshape((X_train.shape[0], X_train.shape[1], 1)), y_train,
                    epochs=10, batch_size=32, validation_data=(X_val.reshape((X_val.shape[0], X_val.shape[1], 1)), y_val))
X_test = scaler.transform(X_test)
y_pred_probabilities = model.predict(X_test.reshape((X_test.shape[0], X_test.shape[1], 1)))
y_pred = np.argmax(y_pred_probabilities, axis=1)

y_test_decoded = label_encoder.inverse_transform(y_test)
y_pred_decoded = label_encoder.inverse_transform(y_pred)
```

Figure 3.8 - Python script to build the prediction model.

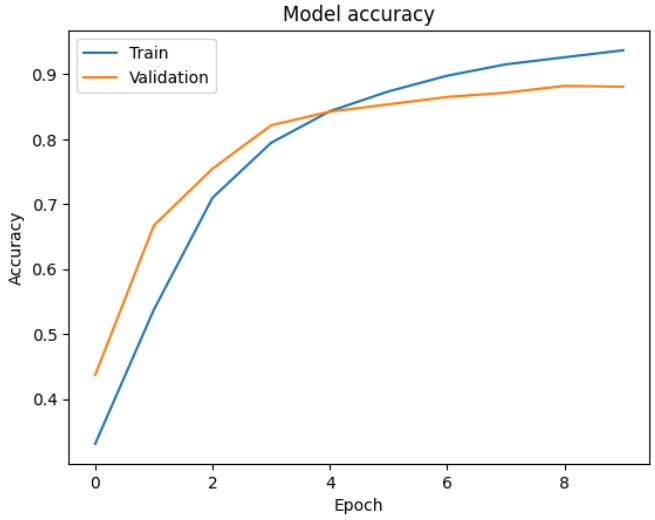

Figure 3.9 - The Model training and validation results per epoch.

We attained a final validation accuracy score of 88%, demonstrating the robustness of our model's performance on unseen data. The slightly higher accuracy rate on the training set (93%) can be explained by the Common Voice dataset's lack of equal representation for certain classes, allowing the model to become more familiar with estimating the age of certain demographics than others.

3.6 Results

In the evaluation phase, the trained model predicted the age category of the speakers remaining in the testing set. Its performance was then assessed using various metrics and techniques, starting with overall performance metrics including accuracy, precision, recall, and F1 Score (see Table 3.4).

Table 3.4 – Performance Metrics Summary.

Metric	Value
Accuracy	0.88
Precision	0.88
Recall	0.88
F1 Score	0.88

Additionally, we also examined the model's performance across individual classes using a confusion matrix (Figure 3.10).

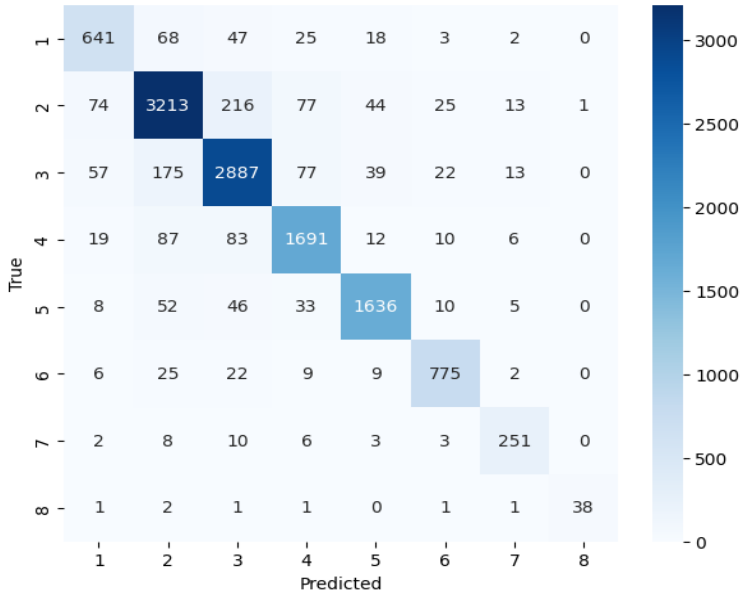

Figure 3.10 - Confusion Matrix for test results.

Moreover, we further explored the model's effectiveness in predicting specific demographics by calculating the accuracy in relation to the gender and accent of the speaker, as shown in Table 3.5.

Table 3.5 – Accuracy by Gender and Accent.

Accuracy	Male	Female	Other
African	0.90	0.86	_
Australia	0.89	0.87	1.00
Bermuda	0.95	0.80	_
Canada	0.89	0.90	1.00
England	0.87	0.89	0.88
Hong Kong	0.75	_	_
Indian	0.90	0.88	_
Ireland	0.90	0.90	0.95
Malaysia	0.86	1.00	_
New Zealand	0.85	0.88	_
Philippines	0.96	0.90	_
Scotland	0.88	0.96	0.75
Singapore	0.90	0.76	_
South Atlantic	0.88	_	_
US	0.87	0.88	0.92
Wales	0.92	_	_

Overall, accuracy varies across different genders and accents, with values generally exceeding 0.85. When gender is not taken into account, the model still performs relatively well.

However, accuracies for Hong Kong, Singapore, and South Atlantic accents are relatively lower, likely due to their limited representation in the dataset.

3.7 Conclusion

In conclusion, after a thorough process from research and planning to realization, we have obtained very good results with our model, although there is always room for improvement. The weak aspect of our study is the lack of equally diverse datasets available, with a more balanced distribution of age, gender, and accent demographics. Additionally, a larger quantity of speech segments can help familiarize the model with more distinct features, thereby improving its overall performance, and although the LSTM structure is effective with sequential data, other DNN types can also be explored for better results.

General Conclusion

Voice-based age estimation is essential in many areas, such as biometric security applications for identification and authentication. The primary objective of this thesis was to explore and develop a robust DNN solution capable of estimating a person’s age from their voice attributes.

By the end of this project, we successfully managed to implement an RNN LSTM model using the Common Voice dataset, focusing on an all-English-speaking demographic across several regional accents. The model was trained to extract relevant spectral features and MFCCs from the voice and utilize gender and accent as additional characteristics to estimate the age of the speakers. Our results show the efficiency and robustness of the model performance within the examined demography, validating the potential of its use in real-world applications.

This endeavor has the potential to revolutionize a wide range of applications, from personalized services to security and healthcare. Reliable voice-based age estimation can optimize user experiences through targeted recommendations, enhance security measures, and contribute valuable insights to medical diagnostics, police investigations, and social sciences studies.

Some of the challenges encountered include restrictions on the model's applicability due to the demographic limitations and disparities within the dataset. Additionally, our computational capacities limited the flexibility to experiment with different DNN structures, such as CNNs, or to use a larger and more comprehensive set of MFCC components without having to resort to a smaller dataset. Furthermore, this being a new application area, there is a notable scarcity of research, especially at the level of master's theses, dedicated to voice-based age estimation.

A potential critique of this study is that voice changes during adulthood can be subtle, making age estimation for older individuals more challenging [10]. This issue is particularly

pronounced in our model, which has been trained on a dataset with a smaller representation percentage for older individuals. Additionally, the high computational requirements of the LSTM structure could slow the system's ability to provide real-time age estimations.

Moving forward, the journey does not end here. As AI technology advances, voice-based age estimation will continue to be explored and refined for better results. Future research should aim to include more regional accents and languages, ensure equal representations across genders and ages groups. Additionally, exploring more complex DNN structures and voice features, despite their higher computational complexity, could enable the estimation of more precise age categories, such as each distinct decade, quinquennial, or even per year.

Bibliography

[1]	National Institute on Deafness and Other Communication Disorders, « How Does the Human Body Produce Voice and Speech?,» 12 April 2022. [En ligne]. Available: https://www.nidcd.nih.gov/sites/default/files/2023-03/how-human-produce-voice-and-speech-2.pdf.
[2]	K. V. R. a. P. T. ELKarazle, «Facial Age Estimation Using Machine Learning Techniques: An Overview,» *Big Data and Cognitive Computing 6*, p. 128, 2022.
[3]	P. G. R. &. K. A. Punyani, «Neural networks for facial age estimation: a survey on recent advances,» *Artificial Intelligence Review 53*, p. 3299–3347 , 2020.
[4]	K. B. a. H. -S. Y. H. -J. Kim, «Age and Gender Classification for a Home-Robot Service,» chez *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, Jeju, Korea (South), 26-29 Aug, 2007.
[5]	M. J. Y. C. S. Anvarjon Tursunov, «Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms,» *Sensors*, p. 5892, 2021.
[6]	H. M. M. T. A. S. F. R. a. M. M. Davood Mahmoodi, «Age Estimation Based on Speech Features and Support Vector Machine,» chez *Conference on Electrical Engineering (CEEC)*, Colchester, UK, 13-14 July 2011.
[7]	D. H. Damian Kwasny, «Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks,» *Sensors vol. 21,14 4785*, July 13, 2021.
[8]	R. G.-P. M. U.-M. M. R.-Z. Héctor A. Sánchez-Hevia, «Age group classification and gender recognition from speech with temporal convolutional neural networks,» *Multimedia Tools and Applications*, p. 3535–3552, January 13, 2022.
[9]	Hammer Team, «The world's first voice-based age verification system launched,» 2022. [En ligne]. Available: https://hammerteam.com/the-worlds-first-voice-based-age-verification-system-launched/.

[10]	Primary Care, «Voice Changes: What Can They Tell You as You Age?,» 30 December 2020. [En ligne]. Available: https://health.clevelandclinic.org/voice-changes-what-can-they-tell-you-as-you-age.
[11]	P. Lieberman, «Primer: Acoustics and Physiology of Human Speech,» 30 June 2018. [En ligne]. Available: https://www.the-scientist.com/primer--acoustics-and-physiology-of-human-speech-64383.
[12]	R. M. K. Shivaji J. Chaudhari, «Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition,» *International Journal of Computer Applications (0975 – 8887)*, p. Volume 117 – No. 17, May 2015 .
[13]	Mozilla, «Common Voice Mozilla,» Common Voice Mozilla, [En ligne]. Available: https://commonvoice.mozilla.org/en/datasets. [Accès le 27 May 2024].
[14]	A. K. A.-H. Ameer A. Badr, «Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks,» *Engineering and Technology Journal*, pp. 129-140, 2021.
[15]	U. E. I. O. I. J. F. E. S. E. C. O. a. E. E. O. Iloanusi, «Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca,» chez *6th IEEE International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Johannesburg, South Africa, November 19-20, 2019.
[16]	T. W. Elan Witkowski, «Artificial intelligence assisted surgery,» *Academic Press*, pp. 179-202, 2020.
[17]	A. K. D. V. C. Y. P. Mohammad Wazid, «Uniting cyber security and machine learning: Advantages, challenges and future research,» *ICT Express*, pp. 313-321, 2022.
[18]	B. K, «Introduction to Deep Neural Networks,» July 2023. [En ligne]. Available: https://www.datacamp.com/tutorial/introduction-to-deep-neural-networks.
[19]	Z.-H. Zhou, «Neural Networks,» chez *Machine Learning*, Singapore, Springer Singapore, 2021, pp. 103-128.

[20]	P. Baheti, «Activation Functions in Neural Networks [12 Types & Use Cases],» 27 May 2021. [En ligne]. Available: https://www.v7labs.com/blog/neural-networks-activation-functions.
[21]	A. D'Agostino, «Introduction to neural networks — weights, biases and activation,» 27 December 2021. [En ligne]. Available: https://medium.com/@theDrewDag/introduction-to-neural-networks-weights-biases-and-activation-270ebf2545aa.
[22]	M. Y. Z. K. T. Y. Wang Qi, «A Comprehensive Survey of Loss Functions in Machine Learning,» *Annals of Data Science*, pp. 187-212, 2022.
[23]	A. Gupta, «Regularization in Machine Learning,» geeksforgeeks, 18 March 2024. [En ligne]. Available: https://www.geeksforgeeks.org/regularization-in-machine-learning/.
[24]	T. P. N. M. R. M. Farhad Shiri, «A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU,» *ArXiv*, 2023.
[25]	S. Hochreiter et J. Schmidhuber, «Long Short-term Memory,» *Neural Computation*, vol. 9, pp. 1735-80, 1997.
[26]	D. Kuhlman, A Python Book: Beginning Python, Advanced Python, and Python Exercises, Platypus Global Media, 2011.
[27]	Python Software Foundation, «About Python,» 20 April 2012. [En ligne]. Available: https://www.python.org/about/.
[28]	P. Jupyter, «Jupyter,» 2024. [En ligne]. Available: https://jupyter.org/.
[29]	Scikit-learn developers, «Scikit-learn,» 2024. [En ligne]. Available: https://scikit-learn.org/stable/.
[30]	«Why TensorFlow,» 10 November 2015. [En ligne]. Available: https://www.tensorflow.org/about.
[31]	«About Keras 3,» 10 February 2024. [En ligne]. Available: https://keras.io/about/.
[32]	B. C. R. D. L. D. P. E. M. M. E. B. a. O. N. McFee, «librosa: Audio and music signal analysis in python.,» chez *Proceedings of the 14th python in science conference*, 2015.

[33]	«ML-model for gender age detection from voice,» 26 November 2021. [En ligne]. Available: https://www.kaggle.com/code/mehmetsametdurgun/ml-model-for-gender-age-detection-from-voice.
[34]	«Age and gender,» 23 March 2023. [En ligne]. Available: https://www.kaggle.com/code/tigerbunny21/age-and-gender/notebook.
[35]	L. Ganji, «One Hot Encoding in Machine Learning,» 21 March 2024. [En ligne]. Available: https://www.geeksforgeeks.org/ml-one-hot-encoding/.
[36]	T. A. Leo Kristopher PIEL, «Speech-Based Identification of Children's Gender and Age with Neural Networks,» chez *Human Language Technologies – The Baltic Perspective*, IOS Press, 2018, pp. 104 - 111.
[37]	J. M. G. J. W. Grey, «Perceptual effects of spectral modifications on musical timbres,» *The Journal of the Acoustical Society of America*, p. 1493–1500, 1978.
[38]	M. Pilanci, «Signal Processing for Machine Learning Lecture 3 Part II,» Stanford University, October 12, 2021.
[39]	musicinformationretrieval.com, «Spectral Features,» 2024. [En ligne]. Available: https://musicinformationretrieval.com/spectral_features.html#:~:text=Spectral%20rolloff%20is%20the%20frequency,feature..