Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Marie Somnea Heng

Early Soft Biometric Voice Recognition

Master's thesis in Information Security
Supervisor: Patrick Bours
Co-supervisor: Matúš Pleva (Technical University of Košice)
May 2023

NTNU
Norwegian University of
Science and Technology

Marie Somnea Heng

Early Soft Biometric Voice Recognition

Master's thesis in Information Security
Supervisor: Patrick Bours
Co-supervisor: Matúš Pleva (Technical University of Košice)
May 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Early Soft Biometric Voice Recognition

Marie Somnea Heng

May 2023

Abstract

Adults who pretend to be children can pose a threat to children by providing their wrong age on communication platforms to approach children online.

Concerning this topic, studies have been conducted to investigate the human voice regarding age classification. In this master's thesis, a training model prototype was used to classify voices into three groups: child, adult, and transitional age group. The inclusion of a transitional age group in the classification helps to consider the diverse stages of individual voice development.

The classification model prototype was trained using the Samrómur dataset. Testing was conducted using a sample from the Common Voice dataset and the "Children speech recording" dataset.

The available information did not include details about the distinction between their labelled verified and non-verified audio files. Therefore, two versions of the Samrómur dataset were created for training the model: one with only verified datasets and another with the complete dataset. The model trained with the verified dataset achieved an accuracy of 95.23%, while the model trained with the complete dataset achieved an accuracy of 90.68%. Both showed signs of an overfitted model either in their loss curve or in the model testing with the other datasets.

Maintaining a high accuracy is crucial for practical applicability. A calculation demonstrated that classifying three pieces of three-second audio theoretically results in a 99% accuracy. Therefore, based on the trained model, the speaker's voice can be classified as early as seven seconds. This calculation considers the trimming method, where each subsequent trim overlaps one second onto the previous piece.

Sammendrang

Voksne som utgir seg for å være barn kan utgjøre en trussel mot barn ved å oppgi feil alder på kommunikasjonsplattformer for å henvende seg til barn på nettet.

For dette emnet er det utført studier der man undersøker den menneskelige stemmen angående aldersklassifisering. I denne masteroppgaven ble en treningsmodellprototype brukt for å klassifisere stemmer i tre grupper: barn, voksen og overgangsalder. Inkluderingen av en overgangsaldersgruppe i klassifiseringen bidrar til å vurdere de ulike stadiene av individuell stemmeutvikling.

Klassifikasjonsmodellprototypen ble trent opp ved hjelp av Samrómur-datasettet. Testingen ble utført ved å bruke et utvalg fra Common Voice-datasettet og datasettet "Children Speech Recording".

Den tilgjengelige informasjonen inkluderte ikke detaljer om skillet mellom deres merkede verifiserte og ikke-verifiserte lydfiler. Derfor ble det laget to versjoner av Samrómur-datasettet for opplæring av modellen: en med kun verifiserte datasett og en annen med hele datasettet. Modellen trent med det verifiserte datasettet oppnådde en nøyaktighet på 95,23%, mens modellen trent med det komplette datasettet oppnådde en nøyaktighet på 90,68%. Begge viste tegn på en overmontert modell enten i tapskurven eller i modelltestingen med de andre datasettene.

Å opprettholde en høy nøyaktighet er avgjørende for praktisk anvendelighet. En beregning viste at klassifisering av tre stykker med tre sekunders lyd, teoretisk sett gir en nøyaktighet på 99%. Derfor, basert på den trente modellen, kan personens stemme klassifiseres så tidlig som i syv sekunder. Denne beregningen tar i betraktning trimmemetoden, der hver påfølgende trim overlapper ett sekund med det forrige stykket.

Preface

This master's thesis was written as the final step to complete my degree in Information Security with the Digital Forensics track at the Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU).

Undertaking this thesis project has allowed me to expand my fundamental knowledge of machine learning within the context of age group classification by voice. It has been a challenging and rewarding experience throughout my master's degree.

I would like to express my deepest gratitude to my internal supervisor, Patrick Bours, who not only facilitated this unique opportunity for me but also enabled me to work on a project related to his Aiba project within the Erasmus+ Traineeship. This 4-month traineeship was completed at the Technical University of Košice (TUKE), Slovakia. I am truly grateful for Patrick's support and guidance throughout this journey.

I would also like to extend my sincere appreciation to my external supervisor from TUKE, Matúš Pleva, for warmly welcoming me as his host student at TUKE and providing invaluable guidance and insights during the course of this research.

I truly appreciate both of my supervisors for their unwavering support, exceptional patience, and constant availability throughout the course of this thesis. Their considerate presence have been invaluable to me.

Lastly, I would like to express my heartfelt thanks to my loving and supportive family, friends, and the incredible new friends I have made during my study abroad experience there at NTNU in Gjøvik, and of course, to myself for never giving up and persevering until the end. Without their unwavering encouragement and belief in me, I would not have reached this point.

Marie Somnea Heng
Košice, Wednesday 31st May, 2023

Contents

Figures

Tables

Abbreviations

List of all abbreviations:

- **(ANN)** Artificial Neural Network
- **(AR)** Autocorrelation Coefficients
- **(ASR)** Automated Speaker Recognition
- **(CNN)** Convolutional Neural Network
- **(DTW)** Dynamic Time Warping
- **(EER)** Equal Error Rate
- **(FMR)** False Match Rate
- **(FN)** False Negative
- **(FNMR)** False Non-Match Rate
- **(FP)** False Positive
- **(FFT)** Fast Fourier Transforming
- **(FFNN)** Feed-Forward Neural Network
- **(FWE)** Formants Wavelet Entropy
- **(FT)** Fourier Transform
- **(GFCC)** Gammatone Frequency Cepstral Coefficient
- **(GMM)** Gaussian Mixture Model
- **(HMM)** Hidden Markov Model
- **(JFA)** Joint Factor Analysis
- **(KNN)** K-Nearest Neighbors
- **(LDA)** Linear Discriminant Analysis
- **(LFCC)** Linear Frequency Cepstral Coefficients
- **(LPC)** Linear Prediction Coding
- **(LR)** Linear Regression
- **(LDC)** Linguistic Data Consortium
- **(LAR)** Log Area Ratios
- **(MAD)** Median Absolute Deviation
- **(MFCC)** Mel-frequency Cepstral Coefficient
- **(MFSC)** Mel-frequency Spectrogram Coefficient
- **(MLR)** Multiple Linear Regression
- **(MyST)** MyST My Science Tutor
- **(NN)** Neural Network
- **(PARCOR)** Partial Correlation Coefficients
- **(PV)** Phase Voder

- **(PDA)** Pitch Detection Algorithm
- **(PSOLA)** Pitch Synchronized Over-Lap-Add
- **(PR)** Polynomial Regression
- **(PSD)** Power Spectrum Density
- **(PLDA)** Probabilistic Linear Discriminant Analysis
- **(PNN)** Probabilistic Neural Network
- **(RF)** Random Forests
- **(RR)** Ridge Regression
- **(RMSE)** Root-Mean-Square Error
- **(SGMM)** Stranded Gaussian Mixture Model
- **(SGM)** Subspace Gaussian Mixture
- **(SVM)** Support vector machine
- **(SVR)** Support Vector Regression
- **(TUKE)** Technical University of Kosice
- **(CFSC)** The children's Filipino speech corpus
- **(TN)** True Negative
- **(TP)** True Positive
- **(VQ)** Vector Quantization
- **(VM)** Virtual Machine
- **(WCC)** Wavelet Cepstral Coefficient
- **(WP)** Wavelet Packets
- **(WT)** Wavelet Transform

Chapter 1

Introduction

This chapter gives an overview of the scope of the thesis project, what it involves, its keywords, the problem description, its purpose and the research questions in order to achieve the wanted results.

1.1 Topics covered by the project

The thesis' topic aims to classify the age group of a speaker. Therefore, the involved topics to cover will be voice, voice audio datasets and machine learning algorithms.

The ability to automatically identify a person's age group based only on their speech features has made voice analysis a useful method for age group classification. Text-independent age group classification is particularly difficult since it is more applicable to real-world situations yet does not rely on specific text cues or speech. Consequently, a model for age group classification would have to rely entirely on the voice audio signal features.

To train and evaluate the age group classification models, a reliable and diverse voice audio dataset is needed. Existing voice datasets with age labels are limited in their coverage of different age groups. Especially, the voice datasets of underaged speakers, as there are more legal and moral considerations to make than when collecting adult voices. Because of this, a variety of spoken audio collections covering a range of age groups will be scanned for this thesis. In addition, to guarantee the performance of the classification model and the dataset, requirements for the dataset will be defined as well.

Machine learning algorithms are effective tools for age group classification that are based on voice analysis. Support Vector Machines (SVMs), K-Nearest Neighbours (KNN) and Neural Networks (NNs) are just a few of the techniques that have been investigated for audio categorization tasks. This thesis will also

provide insight into several research studies related to text-independent age or age group classification. This ought to provide a first idea of what the outcome might be in the performed classification models in the studies.

This thesis attempts to make a contribution to the field of text-independent age group classification by exploring the analysis of voice features, utilizing available voice audio datasets, and using a classification prototype model. The findings and insights gained from this master's thesis research will not only advance the understanding of age-related vocal characteristics but also hold practical implications in the area of voice-based biometrics.

Furthermore, this thesis seeks to pursue the early stage of age group voice classification in order to tailor better interventions. Further details to the purpose of this, will be elaborated in the sections 1.3 and 1.4.

1.2 Keywords

Age group classification, text-independent, voice classification, soft biometrics, audio corpora, dataset review.

1.3 Problem description

Growing numbers of young children have access to the Internet, which contributes to an increase in child abuse cases. Particularly, there is a growing concern over child grooming because children and adults can communicate via publicly accessible online platforms to share sexually explicit messages and media [1]. Adults who pretend to be children can be a danger to potential child victims, as they deceive them into thinking they are interacting with someone their own age. Many researchers are attracted to the automatic approaches to identify grooming conversations as a result of this situation [2].

1.4 Justification, motivation and benefits

The motivation behind this thesis project origins from the Aiba project by Bours, where chat messages are analyzed to determine to which extent the message consists of sexually exploiting intentions with a minor or if it results in a user not being the person he or she claims to be by age [3] [4] [5]. The next step after chatting would be then voice call, for which the master thesis topic "Early soft biometric voice recognition" was set. There, we assume that a predator gets to talk to the victim, which would reveal the voice of the predator as another piece of information to the chats. This helps to have an additional evaluation factor to assess the actual identity of the predator, as well as the safety of the conversation [6].

Identifying or classifying the age based on the human voice with the help of machines has been subject to several research projects so far, as being able to identify those soft biometrics can be helpful in various areas such as forensics to narrow down the list of suspects, call centres for market research purposes or voice quality improvement [6].

1.5 Research questions

As mentioned in the motivation section 1.4, the purpose that this thesis eventually wants is to expose child predators online. The idea is to have a system that tries to determine the age group child or adult of both parties based on their voices when they are on a call. If the system detects that the indicated age does not match the classified age group, it raises an alert to a human moderator, who could be the parent of the potential victim. The moderator is then advised to listen to the conversation and decide if the alert is a false alarm or a genuine concern. If the automatic system would accurately detect a potential predator posing as a child, they can take immediate action to protect the potential victim. This timely intervention could help prevent any harm to the child and initiate further investigations into the potential predator.

Concerning this matter, the research question in this thesis research project is

How early can the speaker's voice be classified as child or adult?

Following additional sub-questions are set to help answer the research question:

- What are the requirements for the dataset?
- What are the age ranges for classifying children and adult voices
- What should be the ideal audio length of the voice for age group classification training?

The set of sub-questions is related to the accuracy of the classification at one point in an utterance, to warn the human moderator to take a closer look into the matter. For that, it will be necessary to implement a prototype to test with the found voice datasets.

1.6 Contributions

The master thesis project attempts to classify the age group based on the human voice, which will lead to the following contributions:

- A review of the found voice audio datasets.
- Age group classification prototype based on the datasets from the datasets.

Chapter 2

Related Works

This chapter covers topics relevant to the thesis project as mentioned in the supporting topics from section 1.1. Insight into the state of the art of similar or relevant studies concerning the age and binary biological sex classification of the human voice will be given here.

To ensure that the biological characteristics of males and females are consistently referred to throughout the thesis, the term sex will be used in accordance with the definition by the World Health Organisation, as gender refers to the socially constructed and identified characteristics of men and women [7].

2.1 Voice

In this section related topics to the human voice will be covered, which consists of the voice characteristics, then voice audio signal features, a voice in biometrics and the challenges in working with voice and voice audio signals. The main focus will be from the audio computational point of view.

2.1.1 Voice Characteristics

There are multiple characteristics that can be heard and examined in each individual's voice. That is due to the human voice being the combination of numerous distinct frequencies produced by the vocal cords [8]:

- vocal speech (loudness, tempo, stability – physical components)
- tonality of speech (intonation–psychological components, i.e. emotions)
- content of speech

Further features to be found in the human voice are:

- phonetic features: used in vowel recognition, which are the fundamental frequency (F0), first four formant frequencies (F1, F2, F3, F4) and the duration [9] [10]
- pitch: emerging vocal cords vibrations [11]

The voice's features can be differentiated into the categories of physiological and behavioural features. A behavioural feature would be like an accent and a physiological feature is e.g. voice pitch [12] [13]. The influence of sound vibrations on the hearing organs is measured subjectively as loudness, which relies on the magnitude and frequency of the vibrations. A subjective indicator of how quickly certain speech segments are pronounced over time is the pace of speech. The most significant words are typically uttered more slowly, and pace and content can be related as each person speaks at their own unique volume and pace [8].

A person's vocabulary is influenced by his or her social and mental environment. When developed in adolescence by the age of about twenty years old, the characteristics of speech, voice, and intonation, as well as the method of speaking, persist throughout their life. They are specific and exclusively inherited features. One can ascertain the distinctive way of an individual's speaking after scrutinizing the various speech constituents [8].

Various frequency spectra can be used to describe how different voices differ in timbre. When describing a complicated sound wave using a spectrogram, the Fourier transform is the mathematical tool used to analyze the frequency spectrum [8].

2.1.2 Voice Audio Signal Features

Voice contains a lot of information, which can make it difficult to work with. Therefore, one of the priorities in working with voices is extracting only the important parts of information to ensure reliability and efficiency for any system [12].

The features that can be extracted from voice audio signal features are the following:

- amplitude [10]
- Mel-frequency Cepstral Coefficients (MFCCs) [10] [14] [15]
- Mel-frequency Spectrogram Coefficients (MFSCs) [10]
- Gammatone Frequency Cepstral Coefficient (GFCC) [15]
- Formants Wavelet Entropy (FWE) [14]
- Wavelet Cepstral Coefficient (WCC) [14]
- Fundamental Frequency F0 [10]
- Formants F1 until F4 [16]
- Autocorrelation Coefficients (AR) [16]
- Partial Correlation Coefficients (PARCOR) [16]
- Log Area Ratios (LAR) [16]
- Mel Energies [16]
- Lyapunov coefficient [16]

The MFCC, a well-known characteristic utilized widely in most voice or speaker detection systems, replicates the frequency response of the human ear. The essential bandwidth frequencies that the human ear detects are taken into account when designing MFCC filters. Both linearly spaced and logarithmically spaced filters are used by MFCC [17].

Another feature extraction approach that exclusively makes use of filters with linear spacing is linear frequency cepstral coefficients (LFCC). For each frequency, LFCC offers the same information. In comparison to MFCC, LFCC employs a greater number of filterbanks in the higher frequency band of speech. The f-ratio, also known as the inter-to-intra class speaker variability ratio, is substantially larger in LFCC than MFCC [17].

In the last decade, speaker recognition techniques that are based on different types of factor analysis i.e. joint factor analysis (JFA), i-vectors, linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (PLDA) produced improved speaker recognition results [17]. Factor analysis is for reducing a large number of variables in a dataset, as well as highlighting structural relationships between the features [18].

The indices for the cycle-to-cycle changes of fundamental frequency and amplitude are jitter and shimmer, respectively. The voice quality is described using these indices. From one cycle to the next, a speaker's voice has a fluctuating frequency. The assessment of voice stability known as jitter is the cycle-to-cycle variation in fundamental frequency. It is a measurement of the vocal fold vibration periodicity reflection [10] [19]. The indicator for voice amplitude disturbance, however, is shimmer. These characteristics, which identify unique voices, offer speaker-specific information [10].

Various frequency spectra can be used to describe how different voices differ in timbre. When describing a complicated sound wave using a spectrogram, the Fourier transform is the mathematical tool used to analyze the frequency spectrum [8].

The pitch frequency, or the frequency of the impulses of the vocal source arising from the vibrations of the vocal cords, is one of the distinguishing characteristics [8]. The pitch analyzes the audio signal's fundamental frequency across time. The Overlap length and Window length are used to divide the audio stream [15]. The frequency of the basic tone is interpreted as the average estimate over a specific interval in this situation since the frequency of oscillations can be broken by variations in the amplitude, frequency, phase, and presence of noise. Speech signals are complex, non-stationary, nonlinear signals with rapidly varying amplitude and frequency characteristics. The most common decomposition techniques used in speech signal processing are the Fourier transform (FT) and wavelet transform (WT), each of which has benefits and drawbacks [8].

The characteristics of the speech signal that correlate to individual voice traits are typically explained using the frequency spectrum of the signal. One of the most often employed feature extraction methods is the Mel-Frequency Cepstral Coefficient (MFCC). MFCC is a filterbank-based methodology created to reproduce the audio frequency perception of the human ear and extracts prosodic or acoustic features [14] [19] [20] [21]. In the scope of speech recognition, MFCC is often used [15]. Figure 2.1 below shows the process of how the MFCC feature is extracted from a speech sequence.

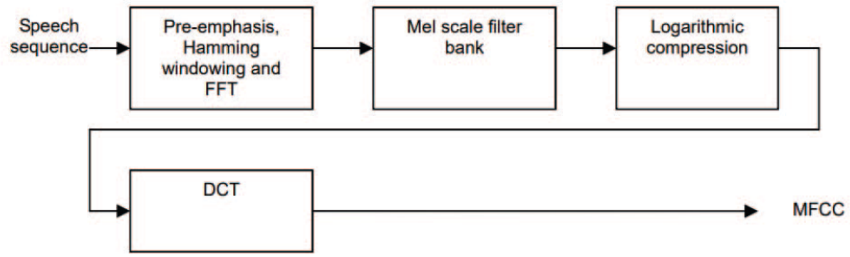


Figure 2.1: Block Diagram for extracting MFCCs feature from input speech segments [15]

In Figure 2.1 it is shown that the input speech sample is first going through a Hamming window and Fast Fourier Transforming (FFT) process, where then after the Mel scale filter bank and logarithmic compression is applied. Lastly, the Discrete Cosine Transform is done from which the MFCC vector is generated [15].

The MFCC can be further computed into Mel-frequency Spectrogram Coefficients (MFSCs) to depict the speech's smoothed spectral envelope. The difference is that the DCT step is not existing [10].

For obtaining the Gammatone Frequency Cepstral Coefficient (GFCC), the speech sample is also going through a Hamming window. Next to the gamma tone filter bank and afterwards logarithmically compressed. The last step is then applying the discrete cosine transform as depicted in Figure 2.2 [10].

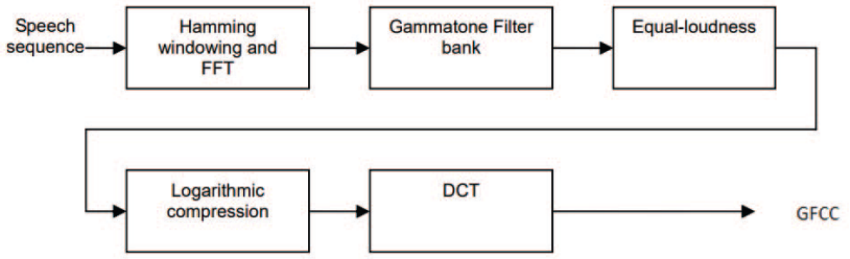


Figure 2.2: Block Diagram for extracting GFCCs feature from input speech segments. [15]

Formants Wavelet Entropy (FWE) has been used in speaker recognition systems and is obtained by computing the formants and the wavelet entropy of the input speech that is filtered. It is applicable to partially obtained voice samples and therefore, often used in forensics. FWE can be applied to vowel-independent as well as vowel-dependent speech, whereas it has been mentioned that the vowel-dependent approach is better [14].

Figure 2.3 shows that the first step in obtaining an FWE feature is recording the speech and filtering. The recording will go through a filter bank to take out unwanted signals from the speech recording. Next is extracting features which consists of two parts. One part is for calculating the formants which incorporate the speaker's vocal tract's acoustic resonance. For that, the Power Spectrum Density (PSD) is applied. There the first five formants are taken for the calculation as they are simple to identify for every human voice. Then the second part is the calculation of the entropies by applying the Wavelet Packets (WP). This part enhances the recognition rate by calculating for every seven nodes of the WP the Shannon entropy [14].

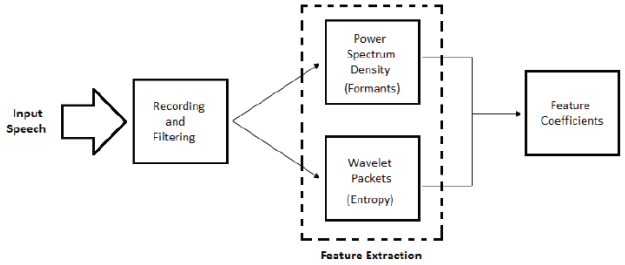


Figure 2.3: FWE Block Diagram [14]

Another way to increase speaker recognition accuracy is the application of Wavelet Cepstral Coefficient (WCC), because of its robustness, which makes it usable alongside fuzzy logic systems in noisy environments [14].

Zbancioc et al. [16] mentioned that in their previous study [22] they were able to extract the features F0, formants F1 to F4, AT, PARCOR, LAR, mel energies and Lyapunov coefficient.

Furthermore, the higher the quality of the voice recordings, the better the results are after processing or training. For that, getting a clear voice out of the recording for processing is required. Unwanted background noises like music or other voices would disturb and lower the quality of the audio. De-noising is a standard procedure for enhancing the audio's quality. In case of having multiple voices in the recording, they will need to be separated [19].

An example of eliminating the noise of clearing one's throat, coughing, teeth clicking and saliva noise. Chabot et al. [23] pointed out how to distinguish them in a spectrogram (see Figure 2.4). As some persons have a tendency to speak at the end of their cough, the spectrogram of the cough signal exhibits harmonics

at the very end of the occurrence. Low-frequency signals with intermittent high-frequency occurrences make up the saliva noise event. More high-frequency material is present during throat clearing, and some harmonics can be heard given how frequently this occurrence is voiced. A distinct impulse with a wide range of frequencies at a specific time is what is heard when teeth click. It is necessary to use a feature extraction technology that can appropriately represent each signal while taking into consideration the significant amplitude and spectral content variance.

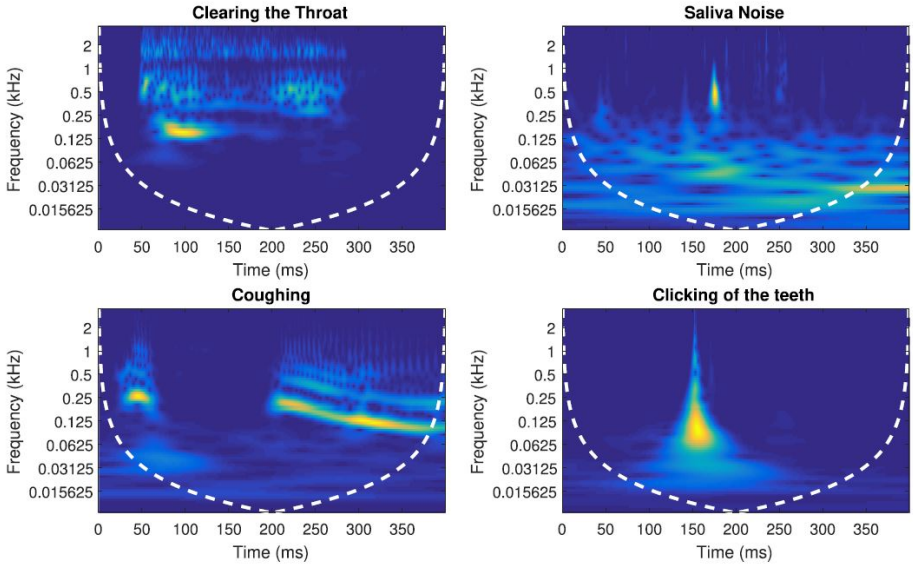


Figure 2.4: Wavelet spectrogram of the audio events "Clearing the Throat", "Saliva Noise", "Coughing" and "Clicking of the teeth" [23]

Kinkiri and Keates [24] mention the following ideal characteristics for features:

- The health (i.e. cold) and age of the speaker should not alter features.
- Features ought to be challenging for others to imitate.
- Noise should not affect the features.

However, in practice, those all can not be covered nor achieved [24].

2.1.3 Classification Models

As mentioned before in subsection 2.1.2, there are several features that can be extracted from a recording. With those features, the following models can be applied for classification in audio processing:

- Vector Quantization (VQ) [14] [25]
- Hidden Markov Model (HMM) [14] [25]
- Gaussian Mixture Model (GMM) [14] [15] [25]
- Pitch Detection Algorithm (PDA) [14]
- Neural Networks (NN) [14]
- Support Vector Machine (SVM) [25]
- Dynamic Time Warping (DTW) [25]
- K-nearest Neighbour (KNN) [15]

Boujnah et al. [25] categorized almost all the techniques mentioned above except for one the PDA. VQ, SVM, and DTW fall into the vector-based approaches. Static approaches are HMM and GMM.

Then a connectionist approach is NN, meaning that it processes elements that are highly interconnected with each other [14] [25].

The VQ is a comparison algorithm, commonly used for speaker recognition, as it can form a set of features to represent the speech data. In other literature, this feature set can also be referred to as a codebook [26] [27]. The advantage of VQ is that it gives out the same results independent from the time sequence of the testing features [26]. However, it can cause inaccuracy due to losing temporal information [14].

The HMM approach provides a statistical depiction of the way a speaker creates sound, which describes the statistical fluctuations of the characteristics. With HMM, the temporal data and underlying speech sounds are well-modelled. Its accuracy is decreased by a number of speakers- and transmission-related factors, and it does not generalize well. As a result, GMM has the advantage over HMM when it comes to text-independent speaker recognition [28].

GMM is a static approach for classification [25]. However, the number of parameters is crucial for a good result, which is why previous work dealt with maintaining good performance while reducing the number of parameters [29]. Other previous works would also propose to extend the GMM due to its large size of parameters that is required, for which e.g. the stranded Gaussian mixture Model (SGMM) or the Subspace Gaussian Mixture (SGM) have been proposed [29] [30].

The PDA uses waveforms created using the autocorrelation approach, where autocorrelation is a correlation between two waves. The PDA minimizes the temporal complexity by cutting the number of comparisons in half and estimates the pitch of an irregular periodic signal [14].

A neural network (NN) is essentially a mechanism for processing information. It is made up of processing components that are closely related to one another. By way of learning, it genuinely helps to tackle pattern recognition issues [14].

There are several approaches for NNs, which are the following:

- Feed-Forward Neural Network (FFNN) [14]
- Probabilistic Neural Network (PNN) [14]
- Convolutional Neural Network (CNN) [31]
- Artificial Neural Networks (ANNs) [32]

Every basic structure of a NN has an input layer, hidden layers and an output layer, where data goes through. How the data is processed in each layer or how the layers are built, depends on the applied approach of NNs [14].

The FFNN is often built of multi-layer nodes. Its name indicates that the data, the NN model is fed with, goes only in one-way (henceforward). Due to the multi-layer nodes, this model takes in multiple data and returns multiple outputs in the output layer [33].

The PNN differentiates from the FFNN by being an unsupervised FFNN model with four layers consisting of the input, pattern, summation and output layer [33].

The CNN is built of two convolutional layers, two fully connected layers, two pooling layers and one softmax layer [31].

The ANN uses the base structure of a NN model but takes the least-squares method to get the weights. The weights are needed to calculate the activation function for output [32].

SVM is frequently employed to categorize human voice recognition and falls under the scope of supervised machine learning. SVM is a non-probabilistic, binary, linear algorithm. Researchers that have conducted speech recognition analyses or regression studies, often employ SVM in the classification process [34] [35] [36].

The Radial Basis Function (RBF) is the most widely used SVM kernel for themes with multi-class classification output because of its high level of accuracy. The class of output in this study is the emotion expressed in human speech [35] [36]. An SVM model is a mapping of the examples as points in space with as much space between the examples of the various categories as possible. On the basis of which side of the gap they fall on, new instances are then mapped into that same space and projected to belong to a category [35]. In a high-dimensional transformed space, the SVM can be used to divide data sets into the best-suited hyperplane that is automatically selected. It exhibits separability as a result of many challenging circumstances [34].

DTW is a sequence matching algorithm that can be found in studies for speech recognition and signal compression for example [37] [38] [39]. DTW has been mentioned to cause a lot of computational costs [39]. DTW belongs to the broad category of algorithms known as dynamic programming. The length of the speech sample and the vocabulary size only have linear effects on the time and space complexity [38].

The KNN's concept is based on the idea that similar observations belong to similar classes. By computing the distances between the unknown object and every object in the training set, KNN determines the k neighbours closest to it [40] [41].

2.1.4 Performance Evaluation Models

There are four terms that are essential for the performance evaluation of a model, which are the True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) values [42].

The true positive (TP) metric is the number of correctly predicted samples as positive. True negative (TN) refers to accurately predicted negative samples. The false negative (FN) metric is the number of samples predicted negatively when they are actually positive. The false positive (FP) metric then refers to samples that are predicted as negative but are actually positive [42].

Table 2.1 depicts the description from before.

Actual Label	**Predicted Label**	**Metrics Definition**
Positive	Positive	True Positive (TP)
Negative	Negative	True Negative (TN)
Positive	Negative	False Negative (FN)
Negative	Positive	False Positive (FP)

Table 2.1: Definition of the terms used to calculate evaluation metrics [42]

The labels assist in computing other factors that are used to evaluate the model's performance and robustness. As using only the accuracy value (see equation 2.1) is not sufficient for measuring the effectiveness and performance of the model, there are the additional measurement factors recall (see equation 2.2), precision (see equation 2.3) and F_1-score (see equation 2.4) are calculated. The accuracy result tells the number of correct predictions of both classes [42].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$F_{\beta_1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.4}$$

2.1.5 Voice Biometrics

By analyzing a person's spoken utterances, speaker or voice recognition systems attempt to identify or verify an identity [13] [21]. The area of speaker recognition is an attractive topic for researchers since it allows applications for voice-based transaction authentication, access control, law enforcement, forensics and many other things [43].

For conducting the identification process, one example would be to match the spoken words with the same phrase that has been stored in text-dependent speaker recognition algorithms. When identifying words, these systems take word dynamics into account. Hidden Markov Models (HMM) and dynamic time warping are the two most popular modeling methods for text-dependent speaker detection (DTW) [21].

For text-independent systems, they do not take feature dynamics into account and instead treat the feature vector as a collection of symbols. In systems of those types, the Gaussian Mixture Model (GMM) or Vector Quantization (VQ) are typically used to model the speakers [21]. The GMM is often used for modeling the speaker data distribution [44]. To build the speaker model and estimate a set of unique parameters, GMM needs a lot of training data (mean, variance, and weights related to each speaker) [21] [44].

Due to the voice's versatility, there have been several approaches to making use of it. The areas it has been used in are biometric personality identification, voice recognition applications, speech or speaker recognition for authentication purposes or to identify soft biometrics [8] [20] [21] [25]. Soft biometrics give information about a person, in which one soft biometric trait would not be sufficient enough to be able to identify one specific person. Those soft biometrics would be eye color, age, or sex for example [45] [46]. The extractable soft biometrics for voice are sex, age, and ethnicity for example [47] [48] [19].

For effective processing of voice datasets, the choice of relevant features that may accurately convey information about a specific speaker's characteristics is according to [8] the most crucial component of successful speaker recognition. The following are the requirements for them that could be also applicable to voice-related experiments, in general [8]:

- the efficiency of presenting information about the features of a particular speaker's speech
- ease of measurement
- stability over time
- practical independence from the acoustic environment
- impervious to imitation

The common way to build a speaker recognition system consists of three steps, which are pre-processing, feature extraction and speaker modelling [49]. This process can be also generally applicable to voice recognition systems [14].

Pre-processing is the starting point of an automated speaker recognition (ASR) model. In order to create an efficient and dynamic ASR system, it is imperative that this procedure is completed on the speech signal input. The voice signal is cleaned in this step of the speaker identification system. The signal is then cleaned up by removing the non-speech parts, which could be silence and non-voice components [14] [49]. Endpoint identification and pre-emphasis are the next preliminary tasks to be finished [49].

Feature extraction is used during the training and testing phases of speech recognition systems. It can be also referred to as front-end pre-processing. It uses feature vector or numerical descriptor sets to transform digital voice signals. The key elements of the speaker's speech are represented in these feature vectors [49].

The last step is speaker modelling, where modelling techniques aim to provide speaker identification algorithms for speech feature matching of the speaker. Speaker models are described as techniques that combine increased speaker-specific information with a compressed volume. State speaker models are created throughout training or enrollment by repeating the specific traits taken from the contemporary speaker. For identification or verification tasks in the recognition state, the speaker model is compared to the current speaker architecture [49].

There is also a standard of voice research at least in the scope of text-independent speaker recognition, the NIST Speaker Recognition Evaluation Standard also known as NIST SRE [44].

Due to the unique physiological nature of the speaker's articulatory apparatus and the specificity of his speech, the speaker's voice and, consequently, the speech signal itself is distinctive [8].

In general, most studies concluded that improvement can be gained in the prediction of those soft biometrics by using deep learning models. To achieve high generalization accuracy rates, these deep learning techniques need a significant amount of training data. However, the collection of extensive biometric datasets as well as soft biometric characteristics like age, sex, and ethnicity may raise privacy and security concerns. Semi-supervised learning is one method for solving this issue. In semi-supervised learning, labeled and unlabeled data are combined to enhance the classifier's generalization capabilities [50][51].

As this paper also aims to identify the sex of the speaker, it may be important to understand the biological setup, development of the human voice and the characteristics of the owner's voice in relation to its sex. This should either help to refine the training and/or classification process or trigger new approaches to classify it, which then would lead to supporting the age classification as well [52].

2.1.6 Challenges in Working with Voice and Voice Audio Signals

In the course of the literature review, some papers raised the issue of the many influences impacting the voice that alter the original voice, which can be put under the term "voice disguises".

There are two kinds of voice disguises, which are technical and natural. Those two kinds are then each further subdivided into deliberate and non-deliberate voice disguise [53].

A deliberate technical disguise typically refers to computer software or an electronic gadget that can change the sound of a speaker. By altering the speaker's basic frequency, the voice can be changed. The most popular methods for achieving this are Pitch Synchronized Over-Lap-Add (PSOLA) or Phase Voder (PV) [53].

In the non-deliberate technological disguise, all distortions and modifications to the voice signal introduced by the communication channel, such as distortions from the recording equipment, band restrictions, applied coding technique, and so on [53].

The unintentional natural voice disguise changes voice parameters brought on by a shift in the speaker's emotional state. A person's voice can change due to a variety of things, including illness (a cold, a sore throat, etc.), alcohol, drugs, and ultimately, the speaker's mental condition, particularly when powerful emotions are present [53].

The intentional imitation of a natural voice can be used to mask the speaker's identity or to imitate another speaker. There is a wide variety of deliberate natural voice masking techniques. A few selected techniques were displayed in a paper by Staroniewicz [53]. They can be broken down into four primary categories: deformation, phonation, phonemic, and prosodic approaches.

Deformation tactics can come across when the vocal tract undergoes forceful physical alterations. The usual deformation techniques include: holding the tongue, pulling the cheeks, pinching the nostrils, or clenching the jaws. The speaker may also place various objects in or over the mouth [53].

Phonation techniques are all vocalization techniques that entail aberrant glottal activity, such as falsetto, trembling voice or whisper [53].

Phonemic methods contain inappropriate allophone use, which happens when a speaker imitates another person or uses a dialect or a foreign accent. The challenges with intonation are addressed through prosodic approaches [53].

Examples of natural prosodic disguise tactics include stress placement, intonation alterations, lengthening or shortening of speech parts, and changing speech pace [53].

Also possibly relevant to the master thesis topic of classifying the actual age group of the speaker is considering the challenges from an adversary's point of view when an adversary imitates a voice whether it is within a technical or skillful means for spoofing attack purposes [54]. By employing phoney biometrics for a real individual, spoofing is accomplished. Fraudsters spoof systems using voice by

using sophisticated speech synthesis, voice imitation or conversion, and recorded playback [15].

Further challenges in speaker identification that again could be relevant to other voice-related experiments are the extraction of informative features, which is speech signal parametrization and finding and building a standard for the dataset [8].

The next subsection will narrow down the scope of voice processing into voice age and age group classification.

2.2 Voice Age Classification

As voice is a continuously changing trait of a human being which makes it possible to estimate the age of a person based on their voice [8] [19]. One of the few researchers in this field demonstrated that as children become older, their vocal characteristics alter, which affects speaker recognition abilities [17]. Additionally, adults have a more heavy voice than children with their high-pitched voice, in which there were reports that age estimation could be possible [55].

Due to the quick evolution of the vocal tract anatomy and the rapid development of the brain, children and teenagers are the most diverse group of speakers with the most distinct differences. The vocal tract doubles at least in size over the first twenty years of life, growing from 8 cm in infants to 18 cm in adults [56].

It is characterized by an extremely steep growth curve that reaches 80% of adult size in infancy, followed by slower, more gradual growth until adulthood. Moreover, distinguished by an earlier, steeper growth curve that reaches 25–40% of adult size, followed by later, consistent growth until puberty [56].

Due to differences in anatomy and motor skills between the sexes, both fundamental frequency and formant frequencies progressively decrease during maturation. Although developmental changes are not linear, they are most noticeable in the early years of life, from infancy to four years of age [56].

The fastest voice change occurs throughout childhood as a result of the larynx, vocal folds, and supporting structures' rapid expansion. The mean fundamental frequency (F0) of females decreases consistently from 225 Hz in ages 20 to 29 to 195 Hz in ages 80 to 90 throughout adulthood. F0 decreases for males until around their 50s. After that, it steadily increases again. A measurement of vocal fold vibration periodicity is called jitter. Some studies found that older women had greater mean jitter values than younger women. Between young and old men, there were also noticeable jitter differences [19].

Identified difficulties in voice recognition, which are background and channel noise, variable and subpar telephones, and extreme hoarseness, fatigue, or vocal stress are made worse by changes in children's voices. For that, there have been active studies towards the extraction of meaningful information from speech over the past three decades [17].

In the following section, recent studies were selected to gain an understanding of how the age classification by voice was performed.

2.2.1 Studies in Voice Age Classification

In a study by Purnapatra et al. [17] they collected voices from 30 subjects over a span of 2.5 years between 4 and 11 years of age at enrollment.

In order to evaluate the performance of longitudinal speaker recognition in children over a period of 2.5 years, state-of-the-art features and algorithms were examined. 20 and 60 coefficients for each of the two feature extraction strategies were tested on the two distinct feature sets MFCC and LFCC. Table 2.2 shows that in comparison to other algorithms and features, the MFCC20 and GMM combination presented the best performance in terms of False Match Rate (FMR) and False Non-Match Rate (FNMR). The Equal error rate (EER) ranged from 22% at a 6-month time instance to 42% at a 30-month time instance. Overall, MFCC and LFCC performed better with 20-dimensional feature vectors than with 60-dimensional features. Also, nearly every feature-algorithm combination failed to perform with an EER between 42% and 56% throughout a 30-month period [17].

Feature	Algorithm	EER 6mo	EER 12mo	EER 18mo	EER 24mo	EER 30mo
MFCC20	GMM	22	26	30	24	42
MFCC20	ISV	48	46	56	52	54
MFCC20	JFA	34	38	35	40	43
MFCC60	GMM	36	38	40	43	42.5
MFCC60	ISV	36	44	40	46	46
MFCC60	JFA	43	37	44	46	52
LFCC20	GMM	26	34	29	40	48
LFCC20	ISV	48	47	50	59	56
LFCC20	JFA	43	38	45	44	50
LFCC60	GMM	38	35	41	45	51
LFCC60	ISV	48	52	46	52	54
LFCC60	JFA	44.5	36	42	52	47.5

Table 2.2: Speaker verification performance - Equal Error Rate in % [17]

In a different study by Ilyas et al. [32] multiple machine learning methods were used, which were Random Forests (RF), SVM, Linear Regression (LR), Polynomial Regression (PR), Ridge Regression (RR) and Artificial Neural Networks (ANNs). To evaluate their performance, 140 participants aged between 6 and 60 years old were invited for an auditory perception test, resulting in a total of 837 completed tests by the participants to classify the voice dataset by age groups. The study's results had an accuracy between 86% and 92% of reasonable classification and 98.2% of age estimation with a root-mean-square error of 2.6 years.

A recent study on automating the estimation of a child's age from voice audio signals was conducted. The acquisition of the voice audio signal data was self-acquired by hiring 255 Czech native children, legally represented by their guardians and with their consent, consisting of 132 girls and 123 boys from the age of 4 to 15 years old. For the study, the children needed to be at least 4 years old, so that they were able to do the study protocol. They were also partitioned into six two-year age subgroups. The study also described the set-up of the voice audio signal recording and what the recording contained [56].

Each utterance was manually evaluated by an expert in acoustic speech with more than ten years of expertise in order to establish the reference values. Following the previously used methodology, the manual evaluation includes a global visual review of the linear prediction coding (LPC). Selected portions with the highest levels of overall format stability and the three formants F1, F2, and F3 visibility were included in the manual analysis. The most stable vowel duration segment of 25–50 ms was used for the estimation, which was done on a steady-state portion of the signal. For the formant analysis, Praat software (version 6.1.09) was used. During the manual estimation, the maximum formant values and the number of formants were changed to produce the most accurate estimate [56].

The Support Vector Regression (SVR) technique produced the best prediction score with a Root-Mean-Square Error (RMSE) of 1.29 and a Median Absolute Deviation (MAD) of 0.20 years. The Multiple Linear Regression (MLR) model achieved the highest predictive performance score across distinct vowels and sexes, with an RMSE of 1.19 and MAD of 0.25 years, followed closely by SVR with the same but less consistent RMSE of 1.19 and MAD of 0.28, both in the boy sample. SVR, however, beat MLR in every other situation, including regression based on single vowel characteristics, females, and mixed sexes [56].

Figure 2.5 displays a confusion matrix for the prediction made using all 20 characteristics and the SVR on the mixed-sex dataset (N = 255). 2600 guesses produced during 100 training and classification trial runs make up the confusion matrix. The matrix's components depict a two-year time window. Findings indicate that with the exception of the youngest group, the majority of forecasts were accurate with regard to age. The algorithm is biased toward older ages when it comes to the youngest group [56].

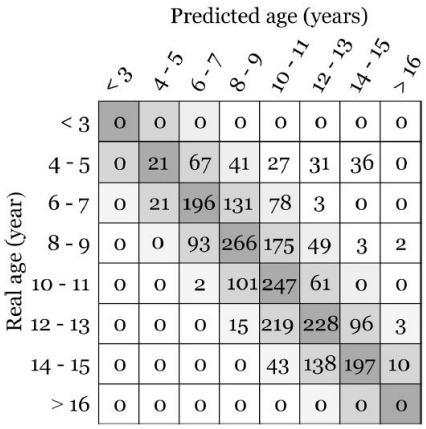


Figure 2.5: Confusion matrix of the 100 runs of the training and testing trials using a mixed-sex dataset and all twenty features in SVR [56]

Chapter 3

Dataset

As mentioned in the Related Works chapter 2 there were studies that either obtained the voice databases from open-source database sites or conducted a voice retrieval with participants.

In this chapter, the acquisition of voice datasets with metadata, as well as the final decision from where to acquire the data for the experiment in this thesis will be explained.

As this master thesis project is done at the host university Technical University of Kosice (TUKE) in the scope of an Erasmus+ Traineeship, an NDA has been signed with the external supervisor at TUKE. This ensures that any obtained corpora through or under the host university TUKE can only be used for this master thesis project.

3.1 Requirements

After the literature review, the set requirements for the dataset were the audio file and relevant metadata containing the biological sex and age. Ideally, data of minors and adults should be of the same corpus. Furthermore, the length of the audio file should be at least 1 second. Any audio file that is longer than 3 seconds will be preprocessed to 3 seconds, which will be described in the section 4.2.

For the age classification, age groups were defined, which resulted in aiming to classify the voice into one of three age groups. Hence group 1 would be voices that are lower than 16 years old. Then group 2 is for voices between 16 and 19 years old and is supposed to serve as a grey zone, due to the difficulty in classifying them as a minor's or adult's voice. Lastly, group 3 contains voices that are older than 19 years old. This grouping will help to achieve the aim of the thesis in determining whether the voice is a minor or an adult.

Regarding the language of the dataset, it was set that it is irrelevant to our project because the focus is on text-independent sex and age classification. In case a platform would provide multiple languages to choose such as Mozilla Common Voice, the choice was primarily English due to its common majoritarian use.

The selected relevant studies from section 2.2.1 showed various discrepancies in the number of participating speakers, processed audio length, gender ratio, age distribution, language and training and testing data ratio. In discussion with the supervisor, specifications of the audio were set based on his expertise.

3.2 Corpora Review

As a starting point, a list of datasets was provided by the supervisor, which was put together into an Excel file and examined. In appendix A the complete documentation with the links to the datasets can be found.

This was a time-intensive process since most dataset platforms did not provide information about whether their dataset also provided metadata which is why the complete dataset needed to be downloaded. The needed metadata were biological sex and age. Another challenge in this search process was to find voice datasets of minors because of limited access and difficulty in obtaining the legal guardian's consent.

The involved people in the following described corpora were contacted where no further information or sample of the dataset could be found. The response rate was low, but since there are corpora of children and adult voices and broad corpora of adult voices were provided and found (see following section 3.3), no further reminders were sent. The received responses were negative in that either they were not available for research distribution, the quality did turn out to not be as needed, or the metadata was not available.

3.2.1 LINDAT/CLARIAH-CZ

The LINDAT/CLARIAH-CZ repository provides the corpus "Speech databases of typical children and children with SLI" (Specific Language Impairment). It also included utterances of 44 children without language impairment, which consisted of 15 boys and 29 girls aged between 4 and 12 years old. Those recordings were conducted between 2003 and 2005 in the Czech language [57].

After downloading the "Healthy.zip" file and checking the audio, it turns out that the majority of the audio files were less than one second. The LINDAT/CLARIAH-CZ corpus was not considered for training the model as it did not provide any metadata. For testing it could be still considerable, as the whole corpus is in the classification age group 1 of lower than 16 years old. However, as mentioned before the set minimum needed length is one second, which makes it in this case not sufficient for testing either.

3.2.2 TalkBank

On the other hand, The TalkBank was a project of Carnegie Mellon University that focused on spoken communication. The project is built on a still accessible database that was created with the assistance of many participants. Over 34 languages were included in an open data-sharing database created from the corpora that were gathered throughout the study. The TalkBank researchers' open-source and free applications enable automatic analysis and searches using a standardized, XML-compatible representation known as CHAT. Among the various databases termed "banks" there are also Child Language Banks. First insights into those banks showed that the corpora did not have a standard, in which some corpora had metadata and some did not or the audio was not necessarily edited. Unedited audio files in the sense that adults' voices from supervising the child were mixed in with the utterance or a whole original 2 hours video file were in there. In addition, the study was active between 1999 and 2004, which makes it likely to be outdated in its quality [58].

Due to the unclear overview of the corpora and not standardized quality and format of the audio files, the corpora in that database will be not suitable for training and only some of the audio files in that database will be suitable for testing.

3.2.3 Children speech recording

The Children speech recording corpus has audio recordings of 11 children from 2016, containing 5 females and 6 males with a median age of 4.9 years old. The content of the recordings is of native and non-native English-speaking children, where each of them had free speech by retelling a picture book and then repeating given sentences [59].

The audio files were provided in full length, as well as trimmed by sentences, which makes it ideal for pre-processing it for testing. For training, it will still not be suitable as the amount of audio files is too little.

For this case, this corpus was selected for testing the trained model if it can classify the children's voices correctly. Further details to this corpus will be discussed in section 3.3.3 under the selected corpora.

3.2.4 CALL-SLT Database

The CALL-SLT Database is a corpus, which was obtained during an experiment in 2015 on school classes in German-speaking Switzerland. 49 students between 14 and 16 years old participated in there. The experiment asked students to play an online CALL game that contained the CALL-SLT, a speech-enabled online CALL tool for beginner learners of English. Therefore, the recordings consist of students solving the exercises in the game containing German and English language [60].

Due to the missing metadata as well as the inconsistent quality of the recordings, they will not be suitable for training. For testing those which do not include the automated voices from the game will usable.

3.2.5 Deeply parent-child vocal interaction dataset

The audio AI company Deeply Inc. captured the conversations between 24 parent-child pairs (a total of 48 speakers) in 2021. An anechoic chamber, a studio apartment, and a dance studio were used for the recordings, and each had a distinct level of reverberation. The purpose of this research was to investigate the impact of the mic's proximity to the source and device. Each experiment was recorded using an iPhone X and a Galaxy S7 at three different distances [61].

The Korean corpus provided the relevant metadata in the form of a JSON file, but the recordings contain the dialogue between the parent and the child, which would make it a time-intensive pre-processing and hence not suitable for training. For testing a few parts of the recordings can be taken that only contains the parent's or the child's voice.

3.2.6 CSTR VCTK Corpus

This CSTR VCTK Corpus includes speech data uttered by 110 English speakers with various accents in 2019. 110 people consisting of 47 males and 63 females between 18 and 38 years old were given 400 sentences to read out loud. The aim of this corpus was to use it for HMM-based text-to-speech synthesis systems. This corpus is also suitable for neural waveform modelling and multi-speaker text-to-speech synthesis systems based on DNN [62].

The quality of the audio was already pre-processed for general use by removing silence and trimming it to one sentence. The audio length seems to be between 1 and 3 seconds making it usable for training and testing.

3.2.7 Samrómur

The Language and Voice Lab (LVL) of Reykjavik University and Almannarómur, the Icelandic Center for Language Technology, collaborated to manage the Samrómur corpus. It has 143,031 (151.8 hours) Icelandic speech recordings, 4,957 of which have been verified. From 2019 until 2022, the recordings were made. The corpus is made up of audio files and a metadata file with the prompts that the participants read. Participants were between 6 and 80+ years [63].

The length of the provided recordings is at least 1 second long, which makes it the ideal corpus for training, as it has metadata and voices of children and adults that covers all the three defined age groups from before in the requirements.

Further details about the Samrómur corpus will be discussed in section 3.3.1 under the selected corpora.

3.2.8 Boulder Learning—MyST Corpus (v0.4.0)

Boulder Learning Inc. created MyST (My Science Tutor) Children's Conversational Speech. It includes around 470 hours of English speech from 1371 students in grades 3-5 speaking with a virtual science instructor about eight different science topics. Between 2008 and 2017, there were two stages of data collection. A total of 227,567 utterances in 10,496 sessions of speech data collection [64].

The corpus is available upon registration at the Linguistic Data Consortium (LDC) and is free of charge. Therefore, this corpus is ideal for adding it to the training model when in the experimenting phase. It was unknown if the metadata is available as it was late by the time the correct site was found for requesting it.

3.2.9 CMU Kids Corpus

The CMU Kids Corpus is a database containing recordings of children reading aloud given sentences in English. It was initially created to produce a training set of child-friendly speech for the SPHINX II automatic speech recognizer for use in the Carnegie Mellon University project LISTEN in 1997. The children were between 6 and 11 years old. There were 52 female speakers compared to 24 males. It has also been described by the authors that, despite the fact that there are more females than boys, they believe that this imbalance should not have much of an impact due to the similar vocal tract lengths of the two at this age. In total, there are 5,180 utterances [65].

The corpus is also available on the LDC site upon registration and purchase which cost between 0$ and 500$. Based on the provided sample, the quality of the corpus seems to be decent and has a length of 22 seconds, which is ideal as it is easier to trim than to concat audios for processing. Therefore, for future purchases this corpus is considerable.

3.2.10 CSLU Kids' Speech Corpus

The CSLU Kids' Speech Version 1.1 is a compilation of 1100 kids in Oregon's Forest Grove School District between Kindergarten and Grade 10 speaking English spontaneously and loudly in 2007. This release includes 1017 files, each of which has 8–10 minutes of speech from each speaker. The original purpose of the corpus was to research the traits of young children's speech at various ages as well as to train and assess recognizers for use in language learning and other interactive tasks involving kids, including training recognizers for deaf kids' language development [66].

The corpus is available on LDC as well and the licence cost between 0$ and 150$. The provided sample indicated a lot of white noise and fidgeting with the microphone, which results in a noisy recording. There were also a lot of silence parts and the audio volume was inconsistent. Therefore, this corpus is not considered for training or testing.

3.2.11 PF-STAR Children's Speech Corpus

The PF-STAR British English children's speech corpus is part of the IST-2001-37599 "PF-STAR: Preparing for Future Multisensorial Interaction Research" EU Framework 5 project in 2006. The corpus was gathered by academics from the Department of Electronic, Electrical, and Computer Engineering at the University of Birmingham in the UK at three locations: a university laboratory and two elementary schools. This corpus includes 158 children's scripted English speech samples, ranging in age from 4 to 14. The recordings are broken down into three sets: a training set (86 speakers, 703 recorded speech files, 7 hours, 29 minutes, and 49 seconds, non-speech included), an evaluation set (12 speakers, 97 recorded speech files, 53 minutes, and 57.579 seconds, non-speech included), and a test set (60 speakers, 510 recorded speech files, 5 hours, 49 minutes, and 47.088 seconds, non-speech included) [67].

Only one paper about the corpus could be found online, in which the author has been contacted. However, there has been no response.

3.2.12 TBALL

The TBALL (Technology Based Assessment of Language and Literacy) corpus was created by researchers from the University of Southern California, the University of California Los Angeles and PPRICE Speech and Language Technology in 2005. The research aimed to validate the impacts of educational technology by linking automatically derived literacy measures from educational technology to later reading performance. 256 children between 5 to 8 years old participated and were from English and/or Spanish native-speaking backgrounds. The given sentences to read aloud were in English. The result is an almost 30,000 speech recording of over 40 hours [68].

One of the TBALL corpus authors was contacted for access to the corpus. Due to the condition of the parent's consent to the data collection, they are not allowed to distribute the data.

3.2.13 CASS_CHILD

The CASS_CHILD corpus is created by the Institute of Linguistics, Chinese Academy of Social Sciences in China from 2009 to 2012. The original purpose of the corpus was to investigate the difference between the Chinese and Indo-European languages, for which 23 Mandarin-speaking children's voices were recorded over a period of time starting from when they were 1 year old until 4 years of age. 13 boys and 10 girls participated, which gave a result of around 570 hours of the recording [69].

The contact details of the researchers were not provided in the paper and online they could not be found either.

3.2.14 Providence Corpus

The Brown University corpus contains longitudinal audio/video recordings from 2002 to 2005 of six English-speaking, monolingual children's conversations with their parents in natural settings. The children ranged in age from 1 to 3 years old. In order to research early phonological and morphological development, the study's goal was to offer a corpus of phonetically transcribed data with connected auditory files. There were 3 boys and 3 girls among the participants. A total of 364 hours of speech make up the corpus [70].

The provided recordings are unedited, which means that they are at their full length per session of around 50 minutes. The recordings also contain the researchers' and parents' voices, as well as background noises. This makes the corpus not considerable for further processing for the training or testing afterward.

3.2.15 Lyon Corpus

The University of Lyon 2 produced the Lyon Corpus between 2002 and 2005. The corpus contains longitudinal audio/video recordings of five French-speaking, monolingual children from the ages of one to three as they engaged in natural household interactions with their mothers. In order to research early phonological and morphological development, the study's goal was to offer a corpus of phonetically transcribed data with connected auditory files. There were 2 boys and 3 girls among the participants. There are 185 hours of speech in the corpus [71].

The quality of the Lyon Corpus is similar to the before mentioned Providence Corpus in subsection 3.2.14, which means full-length and noisy sound in the recordings. This makes the corpus not suitable as well for this project.

3.2.16 Demuth Sesotho Corpus

The Demuth Sesotho Corpus was built between 1980 and 1982 in Lesotho in southern Africa. A longitudinal examination of the linguistic development of four target children throughout the course of their interactions with family members is contained in the corpus. A corpus of 98 hours of speech with roughly 13,250 utterances is the end result. Due to the fact that these data were amassed in impromptu home and neighborhood settings, many of the recordings contain numerous speakers. These include younger peers (ages 2 to 4), older siblings (ages 5 to 7), and adults (adult cousins in their adolescent years, parents, grandparents, and guests). Thus, data from these speakers, who are all listed at the top of each file, can be extracted by researchers interested in studying ordinary Sesotho adult speech. The corpus has roughly 40% of the 4 target children's utterances, 40% of adult utterances, and 20% of peer or older sibling utterances [72].

Again, as in the previously mentioned two corpora in subsection 3.2.14 and 3.2.15, this corpus has the same quality, whereas this one has low volume and low sound quality, making it not suitable for this project either.

3.2.17 CHIEDE

The site where the CHIEDE corpus was provided did not give much information about it. Hence, it is only stated that 59 young participants and 7 hours 53 minutes of recordings make up the CHIEDE corpus of spontaneous child language. Around a third makes up the child's language, with adult speech making up the other two-thirds. The spontaneity of the encounters is CHIEDE's key characteristic and was created in 2005. It was lastly updated in 2008 [73].

The corpus can be obtained by purchasing the license. For academic purposes, this can cost between 100€ and 5000€. A sample was not provided, which made this corpus not considered for future purchases.

3.2.18 TIDIGITS

The TIDIGITS corpus was created and assembled by Texas Instruments, Inc. in 1982. Its goal was to create and test algorithms for connected digit sequence recognition independent of the speaker. There are 326 speakers, each of whom pronounces 77-digit sequences (111 males, 114 women, 50 boys, and 51 girls). Each speaker group is divided into a training and test subset [74].

The corpus is available on the LDC site upon purchase between 0$ and 500$. A sample was not provided, which made this corpus not completely reviewable. Based on the description, the TIDIGITS seems to be a promising corpus if purchasing of corpora is considered for future research.

3.2.19 FAU Aibo Emotion Corpus

The FAU Aibo Emotion Corpus is a collection of children's impromptu, emotionally charged speech recorded while they spoke to Sony's Aibo the pet robot. The corpus is made up of 9 hours of German speech from 51 kids between the ages of 10 and 13 as they converse with Sony's pet robot Aibo. Using syntactic-prosodic criteria, the children's audio recordings were manually divided into brief, syntactically significant "chunks" [75].

Upon further search for contact details to ask for the corpus, it turned out that the researcher who built this corpus for his PhD thesis passed away in 2018.

3.2.20 Swedish NICE Corpus

The Swedish NICE corpus includes spoken exchanges among children between the ages of 8 and 15 who are acting out fairytale characters in a virtual world. The data were collected on a number of times throughout the years 2004–2005. 5,580 utterances from user sound files total were included in the corpus of human-computer communication [76].

Upon corpus request to the authors, there was no response, which did not make it possible to review the Swedish NICE Corpus any further.

3.2.21 SingaKids-Mandarin

The Institute for Infocomm Research and the National University of Singapore created 2016 the SingaKids-Mandarin corpus, which contains 79,843 utterances and 125 hours of data (75 hours of speech) from 255 Singaporean kids between the ages of 7 and 12. All of the speakers could speak at least Mandarin and English and were bilingual or multilingual. This study's objective was to examine Singaporean children's acoustic characteristics [77].

One of the researchers was contacted in regard to the corpus, however, there was a response, in which no review of the corpus could be done.

3.2.22 CFSC

The children's Filipino speech corpus (CFSC) was used in this study to provide offline test data for the evaluation of the RMD system, training data for the generation of speech models, reference speech features (such as pronunciation models and word durations) from good readers, and analysis of actual reading errors discovered in children's Filipino speech. There were two separate recordings made in 2012. The recording was conducted in two parts. About five hours of continuous read speech from 37 pupils, whose ages ranged from about 7 to 11 years, are included in the first section of the CFSC. Twenty boys and seventeen girls make up the group of 37 students. All 37 of the candidates, who attend the University of the Philippines Integrated School in Quezon City, were chosen by their professors as the top readers in their courses. A total of 20 students between the ages of 6 and 9 years contributed nearly three hours of continuous read speech for the second section of the CFSC. There are 9 boys and 11 girls among these students. Students from Makati City's Nemesio Yabut Elementary School make up all 20 contestants [78].

The available age voices would have been of interest for testing and experimenting. However, there was no response from the authors regarding the corpus request, which did not allow any further review of the corpus.

3.2.23 JASMIN Speech Corpus

The JASMIN speech corpus is a collection of Dutch speech from young people, non-native speakers and seniors living in Flanders and the Netherlands. The voice recordings consist of texts read aloud and man-machine dialogues from 2008 and are enriched with various layers of annotation totaling 115 hours of speech. The JASMIN speech corpus is an addition to the Spoken Dutch Corpus [79].

Upon registering to get access to the corpus, the access was still denied. Therefore, no further review of the JASMIN Speech corpus was possible.

3.2.24 Corpus Multilingual Children at Pre-school Age (MEKI)

The Corpus Multilingual Children at Pre-school Age (MEKI) was developed as a component of a study that went along with the rollout of a program to enhance language learning. The youngsters in the groups that were observed ranged in age from eight to twelve. The research was carried out between 2004 and 2006. The objective was to assess the linguistic growth of children between the ages of 5-7 who had not yet started kindergarten. 85 recordings totalling 3 hours, 8 minutes long make up the MEKI corpus version that is archived at the IDS [80].

The corpus could not be obtained due to a denial of accessing it after a few days after registering on the site where the corpus was provided. Hence, no further review could be done.

3.2.25 ALCEBLA

The site, where the ALCEBLA corpus from 2011 is published, gives little information about it. It is mentioned that there are 23 simultaneous bilingual youngsters who live in Germany and attend the first level of the Spanish supplementary school and that the audio recordings are in Spanish. 23 speakers (14 female, 9 male), 66 communications, 64 recordings, and 2122 minutes make up the corpus [81].

The corpus could be obtained on the site of the University of Hamburg after an access request. The corpus did not contain metadata and the audio files are in their full and unedited recorded length. In the recording, the voices of the young participant and the supervising person can be heard. The sound quality is very good, which makes it suitable for testing.

3.2.26 Common Voice

Mozilla's Common Voice is a voice dataset that is freely accessible to the public and is powered by volunteer participants from all around the world. The dataset can be used to train machine learning models for people who want to create speech applications. At the moment it provides corpora of 108 languages [82].

Upon taking a closer look at the English corpora, there are regular updates on the new version of the English corpus. Next to the latest full corpus which is currently 76.39 GB, there is also a corpus segment release of 2.11 GB for version 13.0 for example. Both of them contain metadata in the form of .tsv-files. The recordings are a mix of read sentences or spontaneous talking as the impression gives. Nevertheless, the immense size of this corpus makes it a highly suitable corpus for training just adult voices, since common voice only allows to upload of voices that are from 19-year-old speakers and upwards.

For this case, this corpus was selected for testing if the trained model predicts the adult voices correctly. Further details to this corpus will be discussed in section 3.3.2 under the selected corpora.

3.2.27 aGender

The aGender corpus is an age-annotated database of German telephone voices. A total of 954 paid volunteers spoke for 47 hours of prompted and free text in a manner resembling automated voice services. Male and female participants were equally distributed among the four age cluster groups of children, adolescents, adults, and seniors while choosing the participants. The text primarily comprises of short orders, single phrases, and numbers and was written to be typical of automated voice services. A second database consisted of 659 speakers, 368 of whom were men and 291 of whom were women, who dialed an automated voice portal server and freely responded to one of the two questions [83].

Upon corpus access request by e-mail, one of the authors responded and directed to the Bavarian Archive for Speech Signals where the corpus can be purchased between 455€ and 327€ for scientific purposes. A purchase for this or future works would be not recommended as the quality was low, as the voices were recorded through a phone call.

All those reviewed corpora have been documented in an Excel sheet file. That helped to better judge, which corpora are considerable for training and/or testing. Table 3.1 gives an overview of the reviewed corpora.

After that further self-search was conducted with the Google search engine. It was barely possible to find any free voice dataset of minors and adults of the same corpus for scientific usage. Therefore, no further corpora were found aside from the before-mentioned corpora.

The final selected corpora for the master thesis project will be discussed in the next section.

Corpus	**Audio + Metadata**	**Age range**	**No. of Participants**	**Year**
Lindat	no	4-12	44	2003-2005
TalkBank	no	0.5-18	>100 corpora	1999-2004
Children speech recording	no	unknown	11	2016
CALL-SLT Database	unknown	14-16	49	2015
Deeply parent-child vocal interaction dataset	yes	5-39	823	2021
CSTR VCTK Corpus (v0.92)	yes	18-38	110	2019
Samrómur L2 22.09	yes	5-90	2189	2022
Boulder Learning—MyST Corpus (v0.4.0)	no	unknown	1371	2008-2011; 2013-2018
CMU Kids Corpus	unknown	6-11	76	1997
CSLU Kids' Speech Corpus	unknown	unknown	1100	2007
PF-STAR Children's Speech Corpus	unknown	4-14	ca. 159	2006
TBALL	unknown	5-8	256	2005
CASS_CHILD	unknown	1-4	23	2009-2012
Providence Corpus	yes	1-3	6	2002-2005
Lyon Corpus	yes	1-3	5	2002-2005
Demuth Sesotho Corpus	no	2-4	550	1980-1982
CHIEDE	unknown	unknown	59	2005-2008
TIDIGITS	unknown	unknown	326	1982
FAU Aibo Emotion Corpus	unknown	10-13	51	2009
Swedish NICE Corpus	unknown	8-15	ca. 75	2004-2005
SingaKids-Mandarin	unknown	7-12	255	2016
CFSC	unknown	6-11	57	2012
JASMIN Speech Corpus	unknown	unknown	unknown	2008
Corpus Multilingual Children (MEKI)	unknown	unknown	unknown	2004-2006
ALCEBLA	no	unknown	23	2011
Common Voice	yes	18+	unknown	2023
aGender	yes	7-80	945	2010

Table 3.1: Listing of reviewed corpora

3.3 The Selected Corpora

Among the reviewed corpora there were two corpora available for online download. The Samrómur is chosen for training and testing, while the Mozilla Common Voice dataset is for evaluating the model later on.

3.3.1 Samrómur Corpus

The Samrómur dataset L2 22.09 (available at https://www.openslr.org/130/) is a corpus consisting of voices from children and adults ranging from 5 to 90 years old in the Icelandic language of non-native speakers. 2189 people participated in the collection process between 2019 and 2022.

The corpus consists of 143,031 recordings equivalent to 151.8 hours. After filtering out only the verified ("is_valid"), given "gender" and age out of those recordings 3,552 recordings were suitable for the master thesis project consisting of a total of 213 speakers with a 102 female and 111 male ratio [63]. As there was no further information given on what exactly verified audios were it was decided to include the not verified audios as well. Since not every recorded voice provided information about gender and age, the final complete corpus resulted in 139,640 audio files. For the experiment, there will be one with the verified dataset and the other with the complete dataset to see how much difference it makes between the solely verified and complete dataset.

The files are provided in Free Lossless Audio Codec (FLAC) audio file format, which is around 4 seconds after listening to some randomly selected files.

The provided README.txt file to the corpus mentions that it was executed and collected by the Language and Voice Lab at Reykjavik University in cooperation with Almannarómur, the Icelandic Center for Language Technology. Further, it states that the spoken hour split between female speakers is 101 hours and 28 minutes, while for male speakers 46 hours and 4 minutes are available, which would explain why some speakers have more utterances than others.

3.3.2 Mozilla Common Voice

The English Common Voice Delta Segment 13.0 was downloaded to use for evaluating the training model. Since it is just for evaluation, the latest Common Voice Segment dataset was taken instead of the complete Common Voice Corpus 13.0.

The partial corpus serves as a sample of the complete corpus version. It contains 30,280 English audio files and metadata files.

The metadata files that are outside the audio file folder are of the complete corpus, which is separated into the different groups dev.tsv, invalidated.tsv, other.tsv, reported.tsv, test.tsv, train.tsv and validated.tsv. The provided metadata files for the segmented corpus are invalidated.tsv, other.tsv, reported.tsv and validated.tsv.

Every metadata contains information about the filename, spoken sentence, upvotes, downvotes, age, gender, accents, variant, locale and segment.

The upvotes show the number of people who confirmed that the audio matches the spoken sentence, while the downvotes show the number of people who confirm that the audio does not match the sentence.

The segment column can contain a sentence that belongs to a custom dataset segment.

Randomly selected audio files showed that the files are around 5 seconds long.

3.3.3 Children speech recording

The Children speech recording corpus was downloaded to use for evaluating the model as well since the Common Voice Delta Segment dataset did only contain voice audio files that are from speakers above 18 years old.

The metadata was documented in the folder, which was first divided into free speech or given words and sentences. Then it was divided into files that were cut into sentences or were left at their full length. After that, each folder was encoded with the gender and speaker's background of being a native or non-native English speaker.

For that, the free speech recordings that were cut by sentences were taken into consideration, which was a total of 222 audio files.

Chapter 4

Methodology

This chapter describes the approach and process of setting up the age group classification model, which is structured into the sections preparations, pre-processing, feature extraction, classification model and lastly performance evaluation, where the trained models predict the audio files of the Common Voice dataset.

4.1 Preparations

A single virtual machine (VM) with root access at Openstack (SkyHiGh/stackit) was requested and set up. The request was for an estimated amount of eight CPU cores, 32 GB of RAM, 512 GB of storage space and a Linux operating system, which was granted. The 512 GB were formatted and mounted on /mnt/data. The coding environment was Visual Studio Code by Microsoft. A plug-in tool was installed in Visual Studio Code so that the VM could be accessed via an SSH connection.

The starting point was searching for available Python projects that were related to voice age classification or just voice classification.

The aim of this process was to find an article that would not only provide the code and dataset but also explain the idea and the process behind that.

There were two promising-looking projects. The first one was "Age prediction of a speaker's voice" by Notter [84] and the second one was the "Age Estimation based on Human Voice" project by Arrotta [85] on GitHub.

The first one by Notter gave a helpful explanation and insight into the field of audio data extraction, data cleaning, and feature extraction, as well as on the exploratory data analysis on audio datasets and lastly the machine learning models. The used dataset was the Common Voice from Mozilla. The dataset was downloaded from a provided Kaggle repository, a 14 GB snapshot of the over 70GB original dataset from Mozilla, one of the versions from 2017. The author stated that in the given example around 9,000 audio files were used.

Unfortunately, it was not possible to replicate this elaborated project, due to failures in the feature extraction process with the Samrómur metadata CSV file.

Therefore, the other project by Arrotta was approached, which gave more promising outputs after managing to extract features from the Samrómur metadata and audio file. For this part section 4.2 will elaborate on the process.

Based on that the latter found code was chosen to use it as the base code of this master thesis project. Before the code was further modified and adjusted for the purpose of this thesis, the Samrómur dataset needed some pre-processing, which will be explained in the following section.

4.2 Pre-processing

In regards to the aim of this thesis, which is the early classification of the age group, the audio duration for when the age group should be classified is at three seconds. This duration was taken from the suggested length of the first found project of Notter (see previous section 4.1). For that reason the Samrómur audio files, which are longer than three seconds, needed to be trimmed into 3-second pieces. Those that are less than one second would be ignored. The audio files that were between one second and three seconds were kept, to prevent losing too much data.

To ensure a smooth trim of the audio files, after the first 3-second trim, the next 3-second piece of the audio was trimmed with a one-second shift into the previous piece. If put together, this would create an overlap of the audio pieces.

For example, trimming an eight-second-long audio will return four pieces consisting of three 3-second pieces and one 2-second piece, which is the leftover of the audio. If the leftover piece of the audio was less than one second, then it would be ignored and not saved.

The following figure 4.1 explains visually the trim approach.

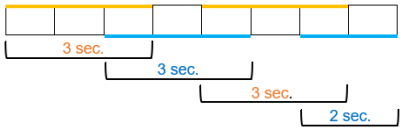


Figure 4.1: General representation of trimming the Samrómur audio files that were over 3 seconds

For editing the audios the pydub Python library [86] was used, which was only able to manipulate audio files of the type WAV, mp3, ogg, flv and other ffmpeg supported files, which did not include audio files of type FLAC. Hence, the filtered-out Samrómur dataset was converted into WAV format with a shell script. At first, the trimming script was tested out on the 3,552 verified audio files. Among them, 2,524 files were over three seconds and needed to be trimmed, which helped to increase the dataset by around 116% with 7,673 verified audio files in total afterwards.

The whole Samrómur dataset was increased by around 89% totalling 264,080 audio samples from 139,640 audio files.

Further audio processing was not needed as there were no disturbing background noises or impairing silences that needed to be removed.

The provided Samrómur corpus metadata was a TSV file, which could be opened in Microsoft Excel for easier filtering. From there, unused columns were removed for a better overview.

The used columns were the filename, sex and age. The sex was encoded into 0 for males and 1 for females. Then a column with the age group was included, which was coded in 0, 1 and 2. 0 for age groups from 0 to 15 years old. 1 for age groups from 16 to 19 years old and 2 for age groups from 20 years old and upwards.

From there the audio files could be then filtered out for the feature extraction process.

4.3 Feature Extraction

For the feature extraction process, the librosa Python package [87] was used. This package is specifically for music and audio analysis.

Here, the following features were extracted [85] [87]:

- Sex and age group: The speaker's sex and age group was taken from the metadata file.
- Spectral centroid: At each frame, the "average" frequency will be calculated, with frequencies being relatively weighted by their energy.
- Spectral bandwidth: It is similar to the centroid but for calculating other moments or variance in the audio.
- Spectral rolloff: The frequency f at which nearly all of the energy in the frame is at frequencies below f will be determined.
- MFCCs: A signal that has a limited number of typically 20 characteristics and each of which accurately captures the general shape of a spectral envelope. It simulates the features of the human voice.

After some little adjustments to the original code to make it executable on the metadata file, it took around six hours to extract the features.

The features were saved into a CSV file containing 26 features, as the MFCC returned 20 characteristics. Table 4.1 shows what the extracted feature outputs of the first five audio files look like.

From there on the next step is the implementation of the classification model.

filename	**gender**	**spectral centroid**	**spectral band-width**	**spectral rolloff**	**mfcc 1**	**...**	**mfcc 20**	**age group**
012159-0319581-3.wav	0	1976.2	1590.4	3420.6	-274.9	...	-1.8	2
012604-0337286-3.wav	0	2437.0	1383.7	3885.4	-274.9	...	1.0	2
012159-0319579-5.wav	0	1114.4	1300.8	2072.0	-270.7	...	-1.0	2
012560-0336635 .wav	0	2351.1	1779.5	4369.1	-405.8	...	3.2	1
010961-0289261-1.wav	1	1498.8	1314.4	2595.8	-283.3	...	2.1	2

Table 4.1: Snippet of extracted features from the 3,225 Samrómur audio files (for presentation purposes numbers have been shortened down to one decimal place and mfcc2-mfcc19 are hidden)

4.4 Classification Model

The used model for classification was the Sequential model from Keras, a Python-written deep learning API that runs on top of the machine learning framework TensorFlow [88]. Tensorflow is an open-source machine learning platform. It has community resources, tools, and libraries.

Tensorflow was created by the Google Brain team, a group of researchers and engineers within Google's Machine Intelligence Research division to undertake machine learning and deep neural network research. The system is broad enough to work in a number of different additional domains as well. At the moment Tensorflow is supported on Python and C++ APIs.

The Sequential model has layers that are stacked linearly and are used in deep learning. The model is basically built with an input layer, hidden layers and an output layer. The input layer consists of the taken-in raw data. Each hidden layer gives the output as the input to the next layer. It is also possible to merge layers, which allows multiple Sequential instances to be blended into a single output. The output layer returns the final prediction or in this case, classification of the model and can be used as the initial level in a brand-new sequential model [89].

The suggested model from Arrotta has an input layer, seven hidden layers and an output layer followed by compiling the model and doing a checkpoint of the model by saving it to a HDF5 file. After that, the training part of the model consisted of setting the epochs and batch size.

An epoch is completing the run through the entire training dataset. The model makes predictions on the full training dataset during an epoch and modifies its parameters in accordance with the estimated loss. Increasing the number of epochs can generally boost a model's accuracy, but it also raises the possibility of overfitting if the model begins to retain the training data [90].

The batch size is the number of training examples used in a single gradient descent iteration during the training of a deep learning model. How many training examples are processed before the model's parameters are changed depends on the batch size. While processing the complete training dataset takes more iterations, smaller batch sizes can result in faster convergence and higher generalization performance. Larger batch sizes can lessen the amount of training dataset processing iterations needed, but they can also cause slower convergence and worse generalization performance [90].

Little adjustments in the code were done in the output layer and in the model compile part.

The units in the output layer were changed to three units instead of eight since there are three classes to be classified by the model, which are the age groups.

In the model compile part the learning rate was added so that the model could take its time to find the minimum loss function.

Since the original code worked with the complete Common Voice dataset, it used 50 epochs and a batch size of 128. This is a high number of datasets to take per batch. For this case with the verified dataset, the best result was with 30 epochs and a batch size of 16, giving an estimated test accuracy of around 93%.

The dropout layers were kept since they help to prevent the model from overfitting during the training process.

With those adjustments figure 4.2 gives a summarised overview of the set-up Sequential model.

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 batch_normalization_27 (Bat  (None, 24)               96
 chNormalization)

 dense_27 (Dense)            (None, 1024)              25600

 batch_normalization_28 (Bat  (None, 1024)             4096
 chNormalization)

 dropout_12 (Dropout)        (None, 1024)              0

 dense_28 (Dense)            (None, 1024)              1049600

 batch_normalization_29 (Bat  (None, 1024)             4096
 chNormalization)

 dense_29 (Dense)            (None, 1024)              1049600

 batch_normalization_30 (Bat  (None, 1024)             4096
 chNormalization)

 dropout_13 (Dropout)        (None, 1024)              0

 dense_30 (Dense)            (None, 1024)              1049600

 batch_normalization_31 (Bat  (None, 1024)             4096
 chNormalization)

 dense_31 (Dense)            (None, 1024)              1049600

 batch_normalization_32 (Bat  (None, 1024)             4096
 chNormalization)

 dropout_14 (Dropout)        (None, 1024)              0

 dense_32 (Dense)            (None, 1024)              1049600

 batch_normalization_33 (Bat  (None, 1024)             4096
 chNormalization)

 dense_33 (Dense)            (None, 1024)              1049600

 batch_normalization_34 (Bat  (None, 1024)             4096
 chNormalization)

 dropout_15 (Dropout)        (None, 1024)              0

 dense_34 (Dense)            (None, 1024)              1049600

 batch_normalization_35 (Bat  (None, 1024)             4096
 chNormalization)

 dense_35 (Dense)            (None, 3)                 3075

=================================================================
Total params: 7,408,739
Trainable params: 7,392,307
Non-trainable params: 16,432
_________________________________________________________________
None
```

Figure 4.2: Model summary

The next step was to find the best split of the dataset for training and testing, which was done by including the stratified k-fold cross-validation [91] in the code. This helps to split the dataset at specific positions that vary for every fold by shifting the position. The stratified instead of the classic k-fold cross-validation was chosen, because it ensures that in addition the dataset is kept in their distribution and proportion in each fold. Figure 4.3 shows how the split dataset works visually.

The split ratio between training and testing is set at 80% training, 10% testing and 10% validation of the dataset, which has been seen as a common ratio split in literature as well in the original code. However, the shuffle variable was set to True so that the random state could be set to 42. The random state can be any number, which needs to be the same when run again, as this helps in terms of the reproducibility of the output. Even if the accuracy and loss results may show differences after each run, it would only be a small difference. The k was set to 10, which means the model is trained 10 times each at different dataset positions. The higher the k-value the more time it takes to run the model, but the more precise the optimal position of splitting the dataset can be found.

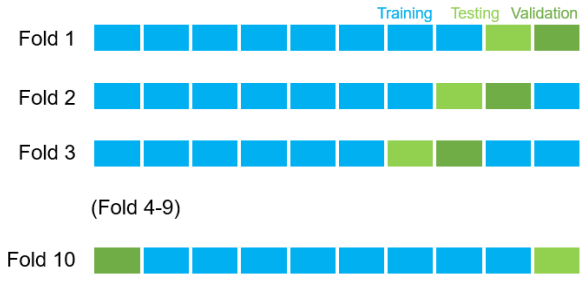


Figure 4.3: Implemented stratified k-folds cross-validation

Again, the verified Samrómur dataset was first trained and then went through testing and validation.

Then the complete Samrómus dataset was run through the model.

Both went through the complete stratified 10-fold cross-validation. This needs to be only run once to find out which fold is the best for each dataset. Taking the best fold means making the most optimal split in the dataset for training, validation and testing.

The result is having one model trained twice with different datasets, one the complete and the other the solely verified Samrómur dataset, which will be used for the next step in the performance evaluation on different datasets.

4.5 Performance Evaluation

Finally, for the performance evaluation the partial corpus version from the Common Voice [82] and the children's speech recording [59] were taken and tested on the model with the function predict from tf.keras.Model [92].

The test audio files were also trimmed the same way as the Samrómur dataset before performing the evaluation. If an audio file was trimmed into multiple parts, it will have its own prediction for each part. Hence, the majority of the occurring predicted age group was set as the prediction output for the complete audio file.

The performance evaluation was performed once with the model that was trained with the verified Samrómur dataset and another once performed with the complete Samrómur dataset.

This gave two results for comparing the two models, which will be presented and discussed in the following chapter.

Chapter 5

Results and Discussions

In this chapter, the results from training the model with the different filtered Samrómur dataset will be presented and discussed.

5.1 Results

5.1.1 Model Training Results

After running the classification model with the verified Samrómur dataset, the following table 5.1 gives the specific output values for each fold.

k	Accuracy	Precision	Recall	F1-score
1	93.75	91.99	91.67	91.83
2	94.01	92.17	92.90	92.52
3	94.53	94.50	93.05	93.71
4	97.91	97.57	97.75	97.65
5	95.30	94.70	94.98	94.84
6	94.26	92.77	91.31	92.00
7	96.08	96.00	95.37	95.66
8	94.78	94.25	93.07	93.57
9	96.08	95.44	96.08	95.75
10	95.56	93.61	93.94	93.77
μ	95.23	94.30	94.01	94.13

Table 5.1: Result of stratified 10-fold cross-validation with the verified Samrómur dataset (in %)

The following table 5.2 shows the result of the model trained with the complete dataset for each fold.

k	Accuracy	Precision	Recall	F1-score
1	89.90	83.53	82.22	82.84
2	91.01	84.25	82.3	83.22
3	91.73	85.32	84.31	84.80
4	90.19	84.75	81.67	83.07
5	91.38	84.90	83.28	84.05
6	91.03	83.22	83.00	83.11
7	89.64	82.73	79.10	80.71
8	91.33	85.07	83.51	84.26
9	90.39	84.07	82.23	83.10
10	90.18	84.09	83.59	83.81
μ	90.68	84.19	82.52	83.30

Table 5.2: Result of stratified 10-fold cross-validation with the complete Samrómur dataset (in %)

To get another point of view on the model's performance during training, the following figure 5.1 shows the accuracy curve of the model trained with the verified dataset. This displays how well the model improves over time. For the loss curve, the figure 5.2 is shown.

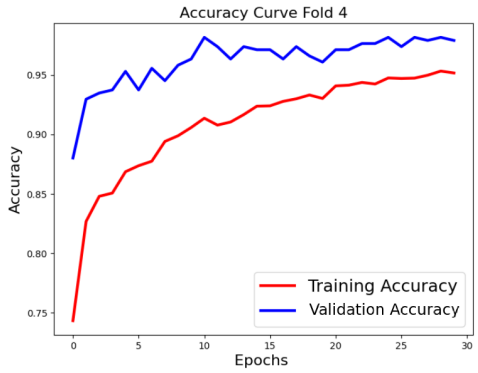


Figure 5.1: Training and validation accuracy curves of the verified Samrómur dataset (4th fold)

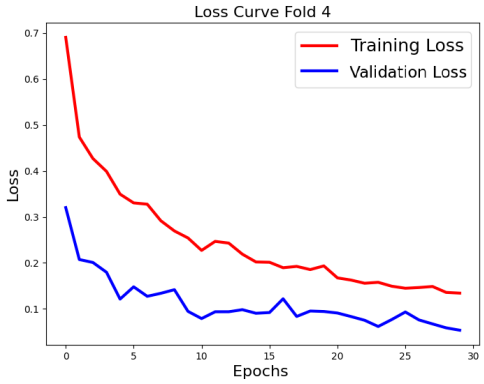


Figure 5.2: Training and validation loss curves of the verified Samrómur dataset (4th fold)

For the trained model with the complete dataset, the accuracy curve is shown in figure 5.3. Its loss curve is displayed in figure 5.4.

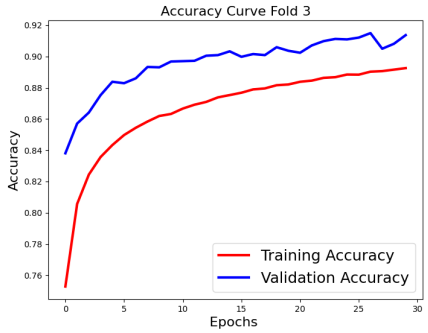


Figure 5.3: Training and validation accuracy curves of the complete Samrómur dataset (3rd fold)

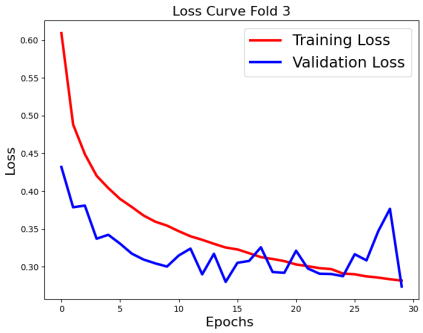


Figure 5.4: Training and validation loss curves of the complete Samrómur dataset (3rd fold)

Furthermore, for analyzing the classification performances of the different trained models, the following figures 5.5 and figure 5.6 show the classification result with the testing and validation dataset in a confusion matrix.

The encoding for age groups on the x- and y-axes in the following confusion matrices are:

- 0: under 16 years
- 1: between 16 and 19 years
- 2: above 19 years

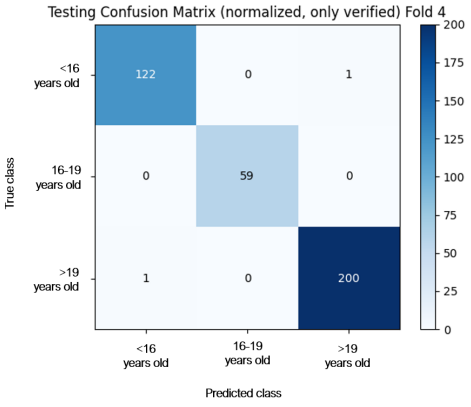


Figure 5.5: Testing Confusion Matrix of the verified Samrómur dataset

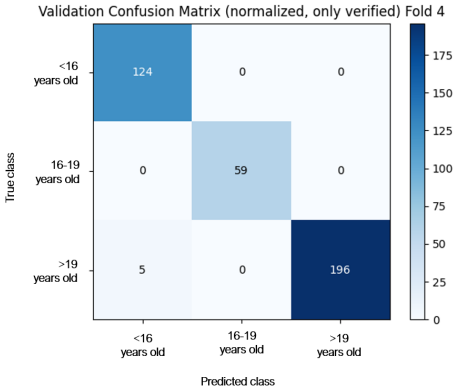


Figure 5.6: Validation Confusion Matrix of the verified Samrómur dataset

The testing and validation confusion matrices with the complete Samrómur dataset are shown in the figures 5.7 and 5.8.

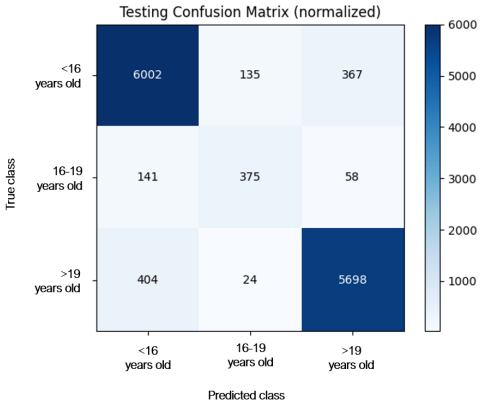


Figure 5.7: Testing Confusion Matrix of the complete Samrómur dataset (3rd fold)

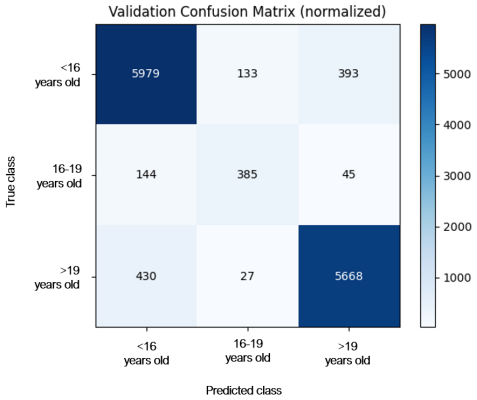


Figure 5.8: Validation Confusion Matrix of the complete Samrómur dataset (3rd fold)

5.1.2 Performance Evaluation Results

After the training with each of the two Samrómur datasets on the model was done, the next step was to perform an evaluation on the model, which is testing it on different children and adult voice audio datasets.

For testing the model on classifying children the "Children speech recording" dataset from subsection 3.3.3 was selected. The age group encoding for all the voices there was set to 0.

The Delta Segment version of the Common Voice English dataset from subsection 3.3.2 was used for testing the model on classifying the adult voices. To support the analysis process, the age groups were encoded according to the original given age group. The reason for maintaining the original age group is to allow a further breakdown in the result and to observe the model's performance more closely.

The age groups encoding the adult voices were:

- 1: teens (from 18 until 19 years old)
- 2: twenties
- 3: thirties
- 4: forties
- 5: fifties
- 6: sixties

Those two datasets were put together as one testing dataset, which was then tested first with the model that was trained with the verified Samrómur dataset and then another time on the model that was trained with the complete Samrómur dataset.

The results are two confusion matrices, where the x-axis is the predicted age groups and the y-axis the true age groups, which are shown in the figures 5.9 and 5.10.

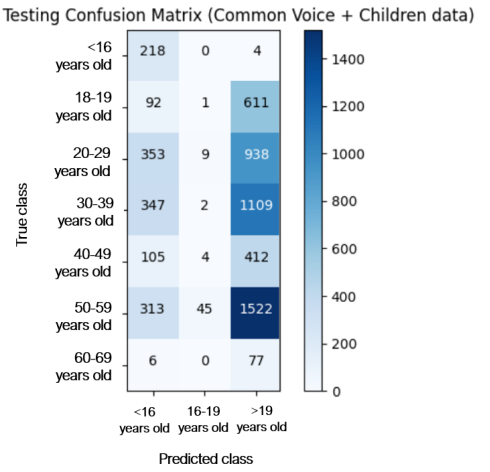


Figure 5.9: Testing Confusion Matrix of the model with a trained verified dataset with Common Voice and Children speech dataset

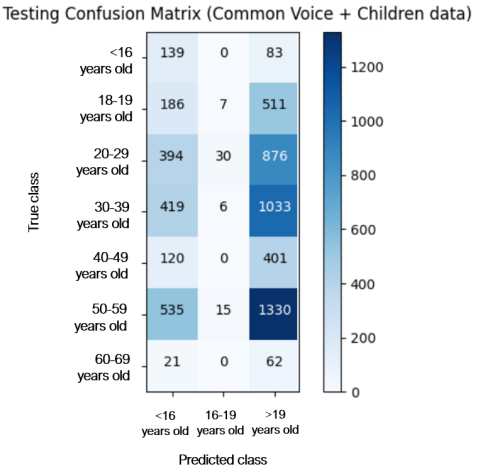


Figure 5.10: Testing Confusion Matrix of the model with a trained complete dataset with Common Voice and Children speech dataset

Now that all the results are presented, the next subsection will analyze and discuss the results.

5.2 Discussions

5.2.1 Model Training Discussion

The various output formats of the result in the previous subsection were needed to observe the model's consistency and classification performance.

With the output of the performance metrics of each iteration of the 10-fold cross-validation, it showed that for the trained model with the verified Samrómur dataset (see 5.1) the average difference between each iteration and performance metric was around 2.13%. Only at the third iteration, it did a sudden improvement in the training, but other than that it maintained a consistent result.

For the complete Samrómur dataset (see 5.2) the average difference was around 1.34%. At the seventh iteration, the model had a sudden fall in its training performance but maintained a consistent result throughout the iterations as well overall.

When comparing the average results between the two tables, the model trained with the verified dataset performed better in accuracy by 4.55%. Then in precision by 10.11% and in the recall by 11.49 %, which resulted in a better F1-score of 10.83%, showing a more balanced precision and recall as well.

By taking a look at figure 5.1 shows that for the verified dataset the training model setup was ideal, as the training and testing accuracy are constantly increasing at a similar percentage, which is identified by the two curves going up alongside each other. For the loss curve in figure 5.2 of the verified dataset, the two curves of training and testing loss can be also seen going down together, which is a positive indication of a well-split dataset and built training model so far.

For the complete dataset, the accuracy curve in figure 5.3 may show a positive result at first sight due to the training and testing curves running close to each other. However, the slow increase by having an almost straight line after around 18 epochs may indicate the first signs of overfitting where the model learns the data too well and does not generalize it anymore. The loss curve in figure 5.4 indicated signs of an overfitted model for the complete dataset since the testing loss increases towards the end, which means that the model is less accurate in its prediction, making the initial suspicion of overfitting based from the accuracy curve valid.

The assumption is that the verified dataset only makes up 2.9% of the complete dataset, which means that 97.1% of not verified data increased the risk of lowering the quality of the dataset significantly.

With the explained results so far, the outputted testing (see figure 5.5) and validation (see figure 5.6) that the trained model with the verified dataset could classify the age group between 16 and 19 years old correctly (group 1), while the other two age groups 0 and 2 each had one misclassification. However, that does not tell that the model can classify age group 1 better than the other two age groups since the age group ratio in the data sample is not evenly distributed.

The confusion matrix of the complete dataset gave a more chaotic result due to the high loss rate, which leads to a higher false classification of the age group than with the verified dataset. As before, this was to be expected due to the already pointed out indications of an overfitted model.

The testing (see figure 5.7) and the validation (see figure 5.8) confusion matrix shows that the age group in the complete dataset is not evenly distributed either, where the ages between 16 and 19 years old (age group 1) make up the minority and therefore more sensitive to the accuracy score in case of misclassification in age group 1.

In this case, it is important to take into account the age group distribution of the dataset right from the beginning of the filtering process. Specifically, the objective is to decrease the size of each age group to match the quantity of the smallest data sample within any age group.

By ensuring an even distribution among the three age groups this should have helped to make the analysis at the end more reliable and comparable.

Although both performed well with an average accuracy of over 90%, it is not a sufficient result to consider the model in practical applications, as that would require at least an accuracy of 99%.

However, it can be calculated with the binomial theorem as shown in equation 5.1 to gain a theoretical outlook on how many pieces it needs to classify to reach an accuracy probability of 99%.

$$Probability = \sum_{k=n+1}^{2n+1} \binom{2n+1}{k} * p^k * (1-p)^{2n+1-k} \tag{5.1}$$

For that, the assumption was that if there was a voice audio file of 11 seconds to classify, there will be five audio pieces after trimming it with the same approach for trimming the trained dataset (see figure).

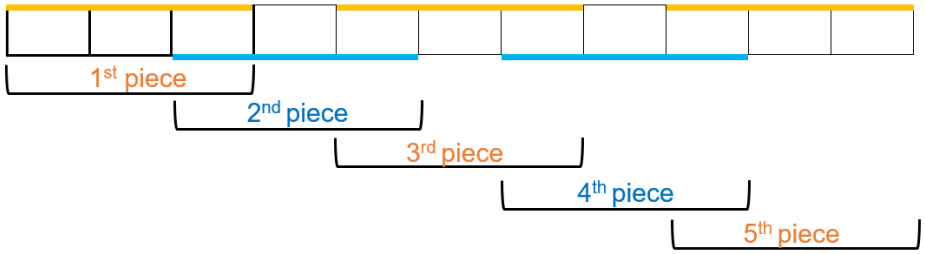


Figure 5.11: Visual presentation of trimming 11-second audio

By starting off with the first calculation of having one out of the five audio pieces correctly classified (see equation 5.2, the result is a probability of 77.38%.

$$Prob. = \binom{5}{5} * p^5 = 1 * 0.95^5 = 0.7738 \tag{5.2}$$

Since the desired 99% is not reached yet, further calculations were done where the correctly classified audio piece is increased by one each time. This led to the result that three audio pieces need to be at least correctly classified to achieve a probability of 99.88% (see equation 5.3 and 5.4.

$$Prob. = \binom{5}{3} * 0.95^3 * (1-0.95)^2 + \binom{5}{4} * 0.95^4 * (1-0.95) + \binom{5}{5} * 0.95^5 \tag{5.3}$$

$$Prob. = 0,0214 + 0.2036 + 0.7738 = 0.9774 \tag{5.4}$$

5.2.2 Performance Evaluation Discussion

Lastly, was the performance evaluation of the classification model by predicting 222 children's voices from the "Children speech recording" 3.3.3 and 5,946 adult voices from the Common Voice dataset 3.3.2.

For a better insight into the classification performance, the age grouping was kept as in the original metadata of the Common Voice, which allowed to point out the weaknesses and strengths of the classification model.

The testing confusion matrix of the model with the trained verified Samrómur dataset in figure 5.9 showed a strong performance in classifying the children's voices, where 98.2% of the children's voices were classified correctly in the children age group.

When taking a look at the adult voice classification, the model still struggles a lot. However, the breakdown of the age groups, showed that the model also has a high performance in classifying the age group 1 and 6. Classifying group 1 as group 2 would be the desired classification for this case anyway since the aim is to differentiate the voice between a child and an adult. The age group 5 would be the next best age group that the model could classify. For the age groups 2, 3 and 4 the model could only classify the majority of them correctly.

The testing confusion matrix in figure 5.10 finally shows the overfitting issue of the model that was trained with the complete Samrómur dataset.

There, throughout all the age groups, the model performed much weaker compared to the previous one. At least 28% of the data samples are classified wrong in each age group, which is rather low for a model with a supposedly average accuracy of 90.68%.

Chapter 6

Conclusion and Future Works

In this chapter, the master thesis will be concluded including findings in the thesis that should be considered in the future.

6.1 Conclusion

Due to the versatility of the voice, researchers have found many ways to make use of it, in which for example based on the human voice the health condition of a patient or the age of the speaker is classified. In this master thesis project, the latter one was taken where the set research question was how early the speaker's voice could be classified as child or adult. As supporting guidance to answer the research question, sub-questions were established in section 1.5).

The first sub-question referred to what the requirements for the dataset are that will be used for training, validating, and testing the classification model.

After conducting a literature and dataset review, the dataset requirements for this project are the audio file and its relevant metadata. The audio file should be at least one second long, and each file should contain only the targeted speaker with no background noise. In the metadata, sex and age are required, as sex is useful as an additional feature in the model training.

The second sub-question regarding what the age ranges for classifying children and adult voices are, was also identified in the dataset requirements set up.

Three age groups were set up, in which the first group consists of under 16-year-olds, the second group consists of ages between 16 and 19 years old and the third group for all voices above 19 years old. The reason for having the second age group is because of the different voice development stages of humans around the puberty period. Hence, this age group serves as a grey zone in this project, where manual classification is rather required.

The last sub-question referred to what the ideal audio length of the voice for age group classification training should be.

The required length of the voice audio was at least one second, where the audios also need to be trimmed if they exceed over three seconds. It all comes down to whether the audio contains enough information that is essential to clearly differentiate the voice between the age groups. Therefore, the ideal audio file length should be three seconds for training the age group classification model in this case.

This leads back to the research question of how early the speaker's voice can be classified as child or adult.

As regards early classification, maintaining an accuracy that allows practical applicability is also important. The implemented classification model trained with the verified Samrómur dataset in this project achieved an accuracy of 95.23% (see 5.1.1). Subsequently, the demonstrated calculation towards the end of section 5.2.1 showed how many pieces need to be classified to reach an accuracy of 99% theoretically. The outcome was that three pieces of each three seconds would be needed in theory to reach the desired accuracy. This means that the answer to the research question is that based on the trained model in this thesis, the speaker's voice can be classified at seven seconds at the earliest. The result is seven and not nine seconds, as the trimming method needs to be considered where every next trim is done with an overlap of one second onto the previous piece.

6.2 Future Works

Further scope to address in the future would be to test the model for weaknesses. That would be for example using fake voices where the speaker pretends a younger voice or using generated children's voices.

Of course, achieved accuracy scores in this project are not enough yet to make it usable in practice, which is why it would be also useful to find other models or to tweak the model for better results and test it more on its robustness. The following suggestions for improving the model are:

- Extracting more features
- Train the model with a balanced dataset ratio of age groups
- Train the model by sex: separate training of male and female voices

Finally, it is recommended to create a dedicated collection of audio recordings for early detection of age groups in online calls. This would help improve the accuracy and reliability of the classification model. Additionally, incorporating sex classification into the model can enhance its effectiveness, as potential predators may falsely claim their sex as well. Another suggestion is to verify the unverified audio samples from the Samrómur corpus to ensure their quality, and then reevaluate the model's performance to identify any differences.

Bibliography

[1] P. Anderson, Z. Zuo, L. Yang and Y. Qu, 'An intelligent online grooming detection system using ai technologies,' in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6. DOI: 10.1109/FUZZ-IEEE.2019.8858973.

[2] M. Ashcroft, L. Kaati and M. Meyer, 'A step towards detecting online grooming – identifying adults pretending to be children,' in *2015 European Intelligence and Security Informatics Conference*, 2015, pp. 1–7. DOI: 10.1109/EISIC.2015.41.

[3] *Aiba*, (Visited 29-05-2023), 2022. [Online]. Available: https://aiba.ai/about-us/.

[4] P. R. Borj, K. Raja and P. Bours, 'Detecting sexual predatory chats by perturbed data and balanced ensembles,' in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021, pp. 1–5. DOI: 10.1109/BIOSIG52210.2021.9548303.

[5] M. A. Fauzi and P. Bours, 'Ensemble method for sexual predators identification in online chats,' in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, 2020, pp. 1–6. DOI: 10.1109/IWBF49977.2020.9107945.

[6] H. Meinedo and I. Trancoso, 'Age and gender detection in the I-DASH project,' *ACM Transactions on Speech and Language Processing (TSLP)*, pp. 1–16, 2011. DOI: 10.1145/1998384.1998387.

[7] Council of Europe, *Sex and gender*, https://www.coe.int/en/web/gender-matters/sex-and-gender, (Visited 21-05-2023).

[8] M. Aliaskar, T. Mazakov, A. Mazakova, S. Jomartova and T. Shormanov, 'Human voice identification based on the detection of fundamental harmonics,' in *2022 IEEE 7th International Energy Conference (ENERGYCON)*, 2022, pp. 1–4. DOI: 10.1109/ENERGYCON53164.2022.9830471.

[9] M. Sharma and K. K. Sarma, 'Dialectal assamese vowel speech detection using acoustic phonetic features, knn and rnn,' in *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, 2015, pp. 1–5. DOI: 10.1109/SPIN.2015.7095270.

[10] S. Soleymani, A. Dabouei, S. M. Iranmanesh, H. Kazemi, J. Dawson and N. M. Nasrabadi, 'Prosodic-enhanced siamese convolutional neural networks for cross-device text-independent speaker verification,' in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, pp. 1–7. DOI: 10.1109/BTAS.2018.8698585.

[11] B. V. Sathe-Pathak and A. R. Panat, 'Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person,' *International Journal of Computer Science Issues (IJCSI)*, pp. 1–4, 2012, ISSN: 1694-0814.

[12] N. Al-Tekreeti and A. A. Ibrahim, 'Speaker voice recognition using a hybrid PSO/fuzzy logic system,' in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1–5. DOI: 10.1109/ISMSIT50672.2020.9254309.

[13] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun and D. Zhang, 'Biometrics recognition using deep learning: A survey,' *Artificial Intelligence Review*, pp. 1–49, 2023.

[14] S. P. Todkar, S. S. Babar, R. U. Ambike, P. B. Suryakar and J. R. Prasad, 'Speaker recognition techniques: A review,' in *2018 3rd International Conference for Convergence in Technology (I2CT)*, 2018, pp. 1–5. DOI: 10.1109/I2CT.2018.8529519.

[15] H. Choudhary, D. Sadhya and V. Patel, 'Automatic speaker verification using gammatone frequency cepstral coefficients,' in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2021, pp. 424–428. DOI: 10.1109/SPIN52536.2021.9566150.

[16] M. D. Zbancioc, R. Butnaru and S. M. Feraru, 'Recognition of voice commands using CNN for romanian language,' in *2022 E-Health and Bioengineering Conference (EHB)*, 2022, pp. 01–04. DOI: 10.1109/EHB55594.2022.9991322.

[17] S. Purnapatra, P. Das, L. Holsopple and S. Schuckers, 'Longitudinal study of voice recognition in children,' in *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020, pp. 1–8.

[18] A. Coni, S. Mellone, M. Colpo, S. Bandinelli and L. Chiari, 'Influence of age and gender on sensor-based functional measures: A factor analysis approach,' in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5054–5057. DOI: 10.1109/EMBC.2015.7319527.

[19] Y. Sun, M. Zhang, Z. Sun and T. Tan, 'Demographic analysis from biometric data: Achievements, challenges, and new frontiers,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 332–351, 2018. DOI: 10.1109/TPAMI.2017.2669035.

[20] A. Boles and P. Rad, 'Voice biometrics: Deep learning-based voiceprint authentication system,' in *2017 12th System of Systems Engineering Conference (SoSE)*, 2017, pp. 1–6. DOI: 10.1109/SYSOSE.2017.7994971.

[21] A. Anand, R. Donida Labati, M. Hanmandlu, V. Piuri and F. Scotti, 'Text-independent speaker recognition for ambient intelligence applications by using information set features,' in *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2017, pp. 30–35. DOI: 10.1109/CIVEMSA.2017.7995297.

[22] M. D. Zbancioc and S. M. Feraru, 'Emotion recognition for romanian language using mfsc images with deep-learning neural networks,' in *2021 International Conference on e-Health and Bioengineering (EHB)*, 2021, pp. 1–4. DOI: 10.1109/EHB52898.2021.9657669.

[23] P. Chabot, R. E. Bouserhal, P. Cardinal and J. Voix, 'Detection and classification of human-produced nonverbal audio events,' *Applied Acoustics*, pp. 1–10, 2021. DOI: doi.org/10.1016/j.apacoust.2020.107643.

[24] S. Kinkiri and S. Keates, 'Speaker identification: Variations of a human voice,' in *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2020, pp. 1–4. DOI: 10.1109/ICACCE49060.2020.9154998.

[25] S. Boujnah, X. Sun, D. Marshall, P. Rosin and M. L. Ammari, '3-step speaker identification approach in degraded conditions,' in *2018 15th International Multi-Conference on Systems, Signals & Devices (SSD)*, 2018, pp. 1–4. DOI: 10.1109/SSD.2018.8570611.

[26] Y. Zhao, X. Zheng, H. Gao and N. Li, 'A speaker recognition system based on vq,' in *2008 3rd IEEE Conference on Industrial Electronics and Applications*, 2008, pp. 1–3. DOI: 10.1109/ICIEA.2008.4582868.

[27] M. R. Fallahzadeh, F. Farokhi, M. Izadian and A. A. Berangi, 'A hybrid reliable algorithm for speaker recognition based on improved DTW and VQ by genetic algorithm in noisy environment,' in *2011 International Conference on Multimedia and Signal Processing*, 2011, pp. 1–5. DOI: 10.1109/CMSP.2011.143.

[28] K. S. Ahmad, A. S. Thosar, J. H. Nirmal and V. S. Pande, 'A unique approach in text independent speaker recognition using mfcc feature sets and probabilistic neural network,' in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, 2015, pp. 1–6. DOI: 10.1109/ICAPR.2015.7050669.

[29] M. Bouallegue, D. Matrouf and G. Linares, 'A simplified subspace gaussian mixture to compact acoustic models for speech recognition,' in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1–4. DOI: 10.1109/ICASSP.2011.5947453.

[30] Y. Zhao and B.-H. Juang, 'Stranded gaussian mixture hidden markov models for robust speech recognition,' in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1–4. DOI: 10.1109/ICASSP.2012.6288870.

[31] X. Chen, Y. Wang, L. Wang and J. Yu, 'Speaker recognition method based on statistical features of spectrograms and CNN,' in *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, 2019, pp. 1–6. DOI: 10.1145/3331453.3361316.

[32] M. Ilyas, A. Othmani and A. Nait-ali, 'Age estimation using sound stimulation as a hidden biometrics approach,' *Hidden Biometrics: When Biometric Security Meets Biomedical Engineering*, pp. 113–125, 2020. DOI: 10.1007/978-981-13-0956-4_7.

[33] K. Daqrouq and T. A. Tutunji, 'Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers,' *Applied Soft Computing*, pp. 231–239, 2015. DOI: https://doi.org/10.1016/j.asoc.2014.11.016.

[34] Li, 'Estimation of intelligibility from received arbitrary speech signals with support vector machine,' in *2005 International Conference on Machine Learning and Cybernetics*, 2005, pp. 1–6. DOI: 10.1109/ICMLC.2005.1527593.

[35] J. Jiang, Z. Wu, M. Xu, J. Jia and L. Cai, 'Comparison of adaptation methods for GMM-SVM based speech emotion recognition,' in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 269–273. DOI: 10.1109/SLT.2012.6424234.

[36] R. A. A., M. Nasrun and C. Setianingsih, 'Human emotion detection with speech recognition using mel-frequency cepstral coefficient and support vector machine,' in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 2021, pp. 1–6. DOI: 10.1109/AIMS52415.2021.9466077.

[37] N. U. Nair and T. Sreenivas, 'Multi pattern dynamic time warping for automatic speech recognition,' in *TENCON 2008 - 2008 IEEE Region 10 Conference*, 2008, pp. 1–6. DOI: 10.1109/TENCON.2008.4766617.

[38] S. Kornsing and J. Srinonchat, 'Exploring dynamic time warping technique to wavelet speech compression,' in *2012 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2012, pp. 1–4. DOI: 10.1109/ECTICon.2012.6254341.

[39] P. Yang, L. Xie, Q. Luan and W. Feng, 'A tighter lower bound estimate for dynamic time warping,' in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 1–5. DOI: 10.1109/ICASSP.2013.6639329.

[40] T.-L. Pao, W.-Y. Liao and Y.-T. Chen, 'Audio-visual speech recognition with weighted KNN-based classification in mandarin database,' in *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, 2007, pp. 39–42. DOI: 10.1109/IIHMSP.2007.4457488.

[41] D. Kamińska and A. Gmerek, 'Automatic identification of bird species: A comparison between kNN and SOM classifiers,' in *2012 Joint Conference New Trends In Audio & Video And Signal Processing: Algorithms, Architectures, Arrangements And Applications (NTAV/SPA)*, 2012, pp. 77–82. DOI: 10.3233/APC220053.

[42] A. Tursunov, Mustaqeem, J. Y. Choeh and S. Kwon, 'Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms,' *Sensors*, pp. 1–19, 2021. DOI: 10.3390/s21175892.

[43] W. T. Hutiri and A. Y. Ding, 'Bias in automated speaker recognition,' Association for Computing Machinery, 2022, pp. 1–18. DOI: 10.1145/3531146.3533089.

[44] K. Sundararajan and D. L. Woodard, 'Deep learning for biometrics: A survey,' *ACM Computing Surveys (CSUR)*, pp. 1–34, 2018. DOI: 10.1145/3190618.

[45] D. Sadhya and S. K. Singh, 'An improved and robust fusion framework for soft biometric traits,' in *2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON)*, 2015, pp. 1–5. DOI: 10.1109/UPCON.2015.7456718.

[46] S. Marrone and C. Sansone, 'An adversarial perturbation approach against CNN-based soft biometrics detection,' in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8. DOI: 10.1109/IJCNN.2019.8851997.

[47] X. Chen, Z. Li, S. Setlur and W. Xu, 'Exploring racial and gender disparities in voice biometrics,' *Scientific Reports*, pp. 1–12, 2022. DOI: 10.1038/s41598-022-06673-y.

[48] A. Dantcheva, P. Elia and A. Ross, 'What else does your biometric data reveal? A survey on soft biometrics,' *IEEE Transactions on Information Forensics and Security*, pp. 1–27, 2016. DOI: 10.1109/TIFS.2015.2480381.

[49] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, 'A survey of speaker recognition: Fundamental theories, recognition methods and opportunities,' *IEEE Access*, pp. 1–28, 2021. DOI: 10.1109/ACCESS.2021.3084299.

[50] A. V. Nadimpalli, N. Reddy, S. Ramachandran and A. Rattani, 'Harnessing unlabeled data to improve generalization of biometric gender and age classifiers,' in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 1–7. DOI: 10.1109/SSCI50451.2021.9660182.

[51] H. J. Galiyawala, M. S. Raval and A. Laddha, 'Person retrieval in surveillance videos using deep soft biometrics,' in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2020, pp. 191–214. DOI: 10.1007/978-3-030-32583-1_9.

[52] F. Chen, R. Togneri, M. Maybery and D. W. Tan, 'Acoustic characterization and machine prediction of perceived masculinity and femininity in adults,' *Speech Communication*, pp. 22–40, 2023. DOI: https://doi.org/10.1016/j.specom.2023.01.002.

[53] P. Staroniewicz, 'Influence of natural voice disguise techniques on automatic speaker recognition,' in *2018 Joint Conference - Acoustics*, 2018, pp. 1–4. DOI: 10.1109/ACOUSTICS.2018.8502372.

[54] H. Guo, Q. Yan, N. Ivanov, Y. Zhu, L. Xiao and E. J. Hunter, 'Supervoice: Text-independent speaker verification using ultrasound energy in human speech,' Association for Computing Machinery, 2022, pp. 1–15. DOI: 10.1145/3488932.3517420.

[55] P. Punyani, R. Gupta and A. Kumar, 'A comparison study of face, gait and speech features for age estimation,' in *Advances in Electronics, Communication and Computing: ETAEERE-2016*, A. Kalam, S. Das and K. Sharma, Eds., 2018, pp. 325–331. DOI: DOI:10.1007/978-981-10-4765-7_34.

[56] M. Novotny, R. Cmejla and T. Tykalova, 'Automated prediction of children's age from voice acoustics,' *Biomedical Signal Processing and Control*, pp. 1–10, 2023. DOI: https://doi.org/10.1016/j.bspc.2022.104490.

[57] *Speech databases of typical children and children with SLI*, (Visited 24-04-2023), 2013. [Online]. Available: https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-1597.

[58] *The TalkBank System*, (Visited 24-04-2023), 2004. [Online]. Available: https://talkbank.org/.

[59] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft and T. Belpaeme, *Children speech recording (english, spontaneous speech + pre-defined sentences)*, 2016. DOI: 10.5281/zenodo.200495.

[60] C. Baur, M. Rayner and N. Tsourakis, *What motivates students to use online call systems? A case study*, Visited 24-04-2023, 2015. [Online]. Available: https://www.researchgate.net/publication/280554791_What_Motivates_Students_to_Use_Online_Call_Systems_A_Case_Study.

[61] D. Inc., *Deeply parent-child vocal interaction dataset*, Visited 24-04-2023, 2021. [Online]. Available: https://www.openslr.org/98/.

[62] *CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)*, Visited 24-04-2023, 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443.

[63] S. Hedström, J. Y. Fong, R. Þórhallsdóttir, D. E. Mollberg, T. Mestrou, S. F. Guðmundsson, Ó. H. Jónsson, S. Þorsteinsdóttir, E. H. Magnúsdóttir, C. L. Richter, R. Pálsson and J. Gudnason, *Samrómur l2 22.09*, Visited 24-04-2023, 2022. [Online]. Available: https://www.openslr.org/130/.

[64] *MyST children's conversational speech*, Visited 24-04-2023, 2021. [Online]. Available: Boulder%20Learning%E2%80%94MyST%20Corpus%20(v0.4.0).

[65] *The CMU kids corpus*, Visited 24-04-2023, 1997. [Online]. Available: https://catalog.ldc.upenn.edu/LDC97S63.

[66] *CSLU: Kids' speech version 1.1*, Visited 24-04-2023, 2007. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2007S18.

[67] M. Russel, *The PF-STAR british english children's speech*, Visited 24-04-2023, 2006. [Online]. Available: http://www.thespeechark.com/pf-star-child-corpus-v1.0.pdf.

[68] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Andersen, S. Narayanan and A. Alwan, 'TBALL data collection: The making of a young children's speech corpus,' 2005, pp. 1–4. DOI: 10.21437/Interspeech.2005-462.

[69] J. Gao, A. Li and Z. Xiong, 'Mandarin multimedia child speech corpus: Cass_child,' 2012, pp. 7–12. DOI: 10.1109/ICSDA.2012.6422462.

[70] *Phonbank english providence corpus*, Visited 24-04-2023, 2002. [Online]. Available: https://phonbank.talkbank.org/access/Eng-NA/Providence.html.

[71] *Phonbank french lyon corpus*, Visited 24-04-2023, 2002. [Online]. Available: https://phonbank.talkbank.org/access/French/Lyon.html.

[72] *CHILDES sesotho demuth corpus*, Visited 24-04-2023, 1980. [Online]. Available: https://childes.talkbank.org/access/Other/Sesotho/Demuth.html.

[73] *CHIEDE corpus: A spontaneous child language corpus of spanish*, Visited 24-04-2023, 2005. [Online]. Available: https://catalogue.elra.info/en-us/repository/browse/ELRA-S0294/.

[74] *TIDIGITS*, Visited 24-04-2023, 1982. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S10.

[75] *FAU aibo emotion corpus*, Visited 24-04-2023, 2009. [Online]. Available: https://www5.cs.fau.de/en/our-team/steidl-stefan/fau-aibo-emotion-corpus/.

[76] L. Bell, J. Boye, J. Gustafson, M. Heldner, A. Lindström and M. Wirén, 'The swedish NICE corpus - spoken dialogues between children and embodied characters in a computer game scenario,' 2005, pp. 1–4. DOI: 10.21437/Interspeech.2005-706.

[77] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma and H. Li, 'SingaKids-Mandarin: Speech corpus of singaporean children speaking mandarin chinese,' 2016, pp. 1–4. DOI: 10.21437/Interspeech.2016-139.

[78] R. M. Pascual and R. C. L. Guevara, 'Developing a children's filipino speech corpus for application in automatic detection of reading miscues and disfluencies,' 2012, pp. 1–6. DOI: 10.1109/TENCON.2012.6412235.

[79] *JASMIN-spraakcorpus*, Visited 25-04-2023, 2008. [Online]. Available: https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus/.

[80] *Multilingual children at pre-school age (MEKI)*, Visited 25-04-2023, 2004. [Online]. Available: https://agd.ids-mannheim.de/MEKI_extern.shtml.

[81] *ALCEBLA*, Visited 25-04-2023, 2020. [Online]. Available: https://www.fdr.uni-hamburg.de/record/1530#.ZFLK-s7P3b1.

[82] *Common Voice Mozilla*, Visited 30-04-2023, 2023. [Online]. Available: https://commonvoice.mozilla.org/en/datasets.

[83] F. Burkhardt, M. Eckert, W. Johannsen and J. Stegmann, 'A database of age and gender annotated telephone speech,' Visited 25-04-2023, 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/262_Paper.pdf.

[84] M. Notter, *Age prediction of a speaker's voice*, Visited 02-05-2023, 2022. [Online]. Available: https://github.com/miykael/miykael.github.io/blob/master/assets/nb/04_audio_data_analysis/nb_audio_eda_and_modeling.ipynb.

[85] L. Arrotta, *Age estimation based on human voice*, Visited 02-05-2023, 2019. [Online]. Available: https://github.com/lucaArrotta/Age-Estimation-based-on-Human-Voice/blob/master/Age%5C%20Estimation%5C%20based%5C%20on%5C%20Human%5C%20Voice.ipynb.

[86] *Pydub 0.25.1*. [Online]. Available: https://pypi.org/project/pydub/ (visited on 10/03/2021).

[87] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen and et al., 'Librosa/librosa: 0.10.0.post2,' 2023. DOI: 10.5281/zenodo.7746972.

[88] *About keras*, Visited 02-05-2023. [Online]. Available: https://keras.io/about/.

[89] R. Deepa, S. Gayathri, P. Chitra, J. J. Jasmine, R. R. Devi and A. Thilagavathy, 'An enhanced machine learning technique for text detection using keras sequential model,' in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, 2023, pp. 1–5. DOI: 10.1109/ICEARS56392.2023.10085174.

[90] J. Brownlee, *Difference between a batch and an epoch in a neural network*, https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/, Visited 02-05-2023, 2022.

[91] *Sklearn model - stratifiedkfold*, Visited 02-05-2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html.

[92] *Keras model*, Visited 02-05-2023. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/Model#predict.

Appendix A

Dataset Review

The following tables are the complete documentation with links to the reviewed datasets from chapter 3.

Provided by	ID	Database/Corpus	contains audio and needed metadata	<18 y/o?	Age from (years)	Age until (years)	Number of participants	Male	Female	Other or not stated sex	Year of Dataset	Year of Paper	Language	Link	Notes
KTU IAESTE Write Up	1	Lindat	no, metadata missing and audio length <2s	yes	4	12	44	15	29	unknown	2003-2005	unknown	Czech	https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-1597	audio is less than 2 secs, wav, 44.1 kHz, 16-bit resolution, mono mode
KTU IAESTE Write Up	2	TalkBank (CHILDES, PhonBank, HomeBank)	yes, but see notes	yes	0.5	18	unknown; maybe 100+ different corpora	unknown	unknown	unknown	1999-2004	unknown	various	https://talkbank.org/	no clear overview, metadata and audio file not assignable dependent on author, only children's voices, some mixed with adult voices (parents), but metadata soley based on children children voice databank: https://childes.talkbank.org/access/ https://phonbank.talkbank.org/access/ https://homebank.talkbank.org/access/
KTU IAESTE Write Up	3	Children speech recording (English, spontaneous speech + pre-defined sentences)	no, age is unknown	yes	unknown	unknown	11	6	5	unknown	2016	unknown	English native and non-native	https://zenodo.org/record/200495#.ZAhq2x-ZNhE	age unknown, only states that the median age is 4.9 years old
KTU IAESTE Write Up	4	CALL-SLT Database	no, metadata missing	yes	14	16	49	unknown	unknown	unknown	2015	unknown	unknown	https://www.researchgate.net/publication/280554791_What_Motivates_Students_to_Use_Online_Call_Systems_A_Case_Study	nikolas.tsourakis@unige.ch -> deliverable emmanuel.rayner@unige.ch Claudia Baur <claudia.baur@bluewin.ch>: Response: Dataset: https://regulus.unige.ch/spokencallsharedtask_3rdedition/ Due to compliance we had to anonymize the data. You can find general information on the participants in my PhD thesis, however this information is not linked to the speech data.
KTU IAESTE Write Up	5	Deeply parent-child vocal interaction dataset	yes, but see notes	yes	5	39	823	unknown	unknown	unknown	2021	unknown	Korean	https://www.openslr.org/98/	all ages are stated as 39 and some 5, parent and child voice not seperated, 16 kHz, 16-bit, mono
KTU IAESTE Write Up	6	CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)	yes	no	18	38	110	47	63	unknown	unknown	2019	English	https://datashare.ed.ac.uk/handle/10283/3443	48 kHz, 16-bits
Matus Pleva	7	Samrómur L2 22.09	yes	yes	5	90	2189	29,90%	67,90%	2,20%	2022	unknown	Icelandic	https://www.openslr.org/130/	meets requirements
Wikipedia - List children's speech corpora:	8	Boulder Learning—MyST Corpus (v0.4.0)	no	yes	unknown	unknown	1371	unknown	unknown	unknown	2008-2011; 2013-2018	unknown	English	https://catalog.ldc.upenn.edu/LDC2021S05	Matus requested
Wikipedia - List children's speech corpora	9	CMU Kids Corpus	unknown, need access rights	yes	6	11	76	24	52	unknown	1997	unknown	English	https://catalog.ldc.upenn.edu/LDC97S63	Fees $0.00 1997 Member $500.00 Non-Member $250.00 Reduced-License $0.00 Extra Copy
Wikipedia - List children's speech corpora	10	CSLU Kids' Speech Corpus	unknown, need access rights	yes	unknown	unknown	1100	unknown	unknown	unknown	2007	unknown	English	https://catalog.ldc.upenn.edu/LDC2007S18	kindergarten to grade 10 at 16kHz; Fees $0.00 2007 Member $150.00 Non-Member $150.00 Reduced-License
Wikipedia - List children's speech corpora	11	PF-STAR Children's Speech Corpus	own, need access	yes	4	14	ca. 159	ca. 82	ca. 81	unknown	2006	2006	British English (given sentences)	http://www.thespeechark.com/pf-star-child-corpus-v1.0.pdf	Martin Russell m.j.russell@bham.ac.uk https://www.birmingham.ac.uk/staff/profiles/computer-science/honorary-staff/russell-martin.aspx

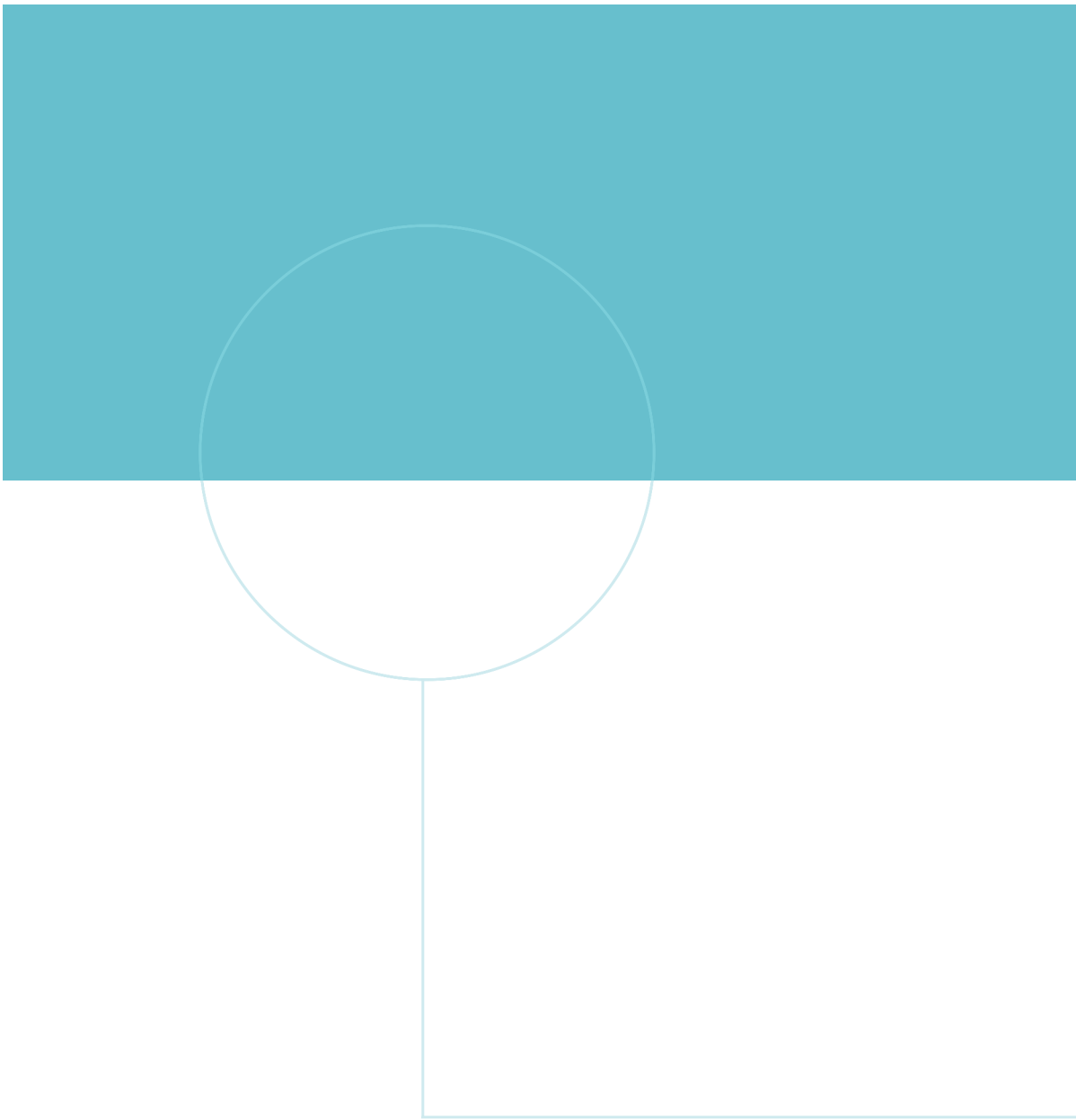
downloaded and most ideal corpus
downloaded for testing
considerable
requested and pending

Provided by	ID	Database/Corpus	contains audio and needed metadata	<18 y/o?	Age from (years)	Age until (years)	Number of participants	Male	Female	Other or not stated sex	Year of Dataset	Year of Paper	Language	Link	Notes
Wikipedia - List children's speech corpora	12	TBALL	er for more informat	yes	5	8	256	ca. 128	ca. 128	unknown	unknown	2005	English non-native	https://www.researchgate.net/publication/221487782_TBALL_data_collection_the_making_of_a_young_children%27s_speech_corpus	Abe Kazemzadeh KAZE7539@stthomas.edu https://software.stthomas.edu/about/faculty-staff/biography/abe-kazemzadeh/ response: Unfortunately, I don't have the data and we were bound by the parent's consent given at the time of the data collection, which didn't allow distributing the data.
Wikipedia - List children's speech corpora	13	CASS_CHILD	unknown, would need to ask author	yes	1	4	23	13	10	unknown	2009-2012	2012	Mandarin	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6422462	
Wikipedia - List children's speech corpora	14	Providence Corpus	yes	yes	1	3	6	3	3	unknown	2002-2005	unknown	English	https://phonbank.talkbank.org/access/Eng-NA/Providence.html	don't know how to read the age range in the description (matches the audio file names at least); has background noise
Wikipedia - List children's speech corpora	15	Lyon Corpus	yes	yes	1	3	5	2	3	unknown	2002-2005	unknown	French	https://phonbank.talkbank.org/access/French/Lyon.html	has adult voices in it; not seperated
Wikipedia - List children's speech corpora	16	Demuth Sesotho Corpus	no	yes	2	4	members, whos	unknown	unknown	unknown	1980-1982	unknown	Sesotho	https://childes.talkbank.org/access/Other/Sesotho/Demuth.html	
Wikipedia - List children's speech corpora	17	CHIEDE	unknown, see notes	yes	unknown	unknown	59	unknown	unknown	unknown	2005-2008	unknown	Spanish	https://catalogue.elra.info/en-us/repository/browse/ELRA-S0294/	100-200€ licence; does not have samples
Wikipedia - List children's speech corpora	18	TIDIGITS	unknown, see notes	yes	unknown	unknown	326	161	165	unknown	1982	unknown	English	https://catalog.ldc.upenn.edu/LDC93S10	contains adult and minor voices at 20kHz; Fees $0.00 1993 Member $500.00 Non-Member $250.00 Reduced-License $0.00 Extra Copy
Wikipedia - List children's speech corpora	19	FAU Aibo Emotion Corpus	unknown, does not link to dataset, just infos	yes	10	13	51	21	30	unknown	unknown	2009	German	https://www5.cs.fau.de/en/our-team/steidl-stefan/fau-aibo-emotion-corpus/	author passed away in 2018
Wikipedia - List children's speech corpora	20	Swedish NICE Corpus	unknown, cannot find dataset in found link	yes	8	15	ca. 75	unknown	unknown	unknown	2004-2005	2005	Swedish	https://www.researchgate.net/publication/221487782_The_Swedish_NICE_corpus_—_spoken_dialogues_between_children_and_embodied_characters_in_a_computer_game_scenario	Johan Boye jboye@kth.se Joakim Gustafson jkgu@kth.se

	downloaded and most ideal corpus
	downloaded for testing
	considerable
	requested and pending

Provided by	ID	Database/Corpus	contains audio and needed metadata	<18 y/o?	Age from (years)	Age until (years)	Number of participants	Male	Female	Other or not stated sex	Year of Dataset	Year of Paper	Language	Link	Notes
Wikipedia - List children's speech corpora	21	SingaKids-Mandarin	unknown, cannot find corpus	yes	7	12	255	unknown	unknown	unknown	unknown	2016	Mandarin	https://www.isca-speech.org/archive_v2/archive_papers/interspeech_2016/pdfs/0136.PDF	nfychen@i2r.a-star.edu.sg rtong@vt.edu
Wikipedia - List children's speech corpora	22	CFSC	unknown, cannot find corpus	yes	6	11	57	29	28	unknown	unknown	2012	Filipino	https://ieeexplore.ieee.org/document/6472225	ronaldmpascual@gmail.com rowena.guevara@eee.upd.edu.ph
Google search: speech children corpora		CLARIN spoken corpora collection	unknown	unknown	unknown	unknown	unknown	unknown	unknown	unknown	unknown	unknown	unknown	https://www.clarin.eu/resource-families/spoken-corpora	keyword search within website with "child" -> 4 results listed in the following rows
CLARIN spoken corpora collection	23	JASMIN Speech Corpus	unknown	yes	unknown	unknown	unknown	unknown	unknown	unknown	2008	unknown	unknown	https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus/	would have minor and adult voices; still could not download after registering with tuke account
Google search: speech children corpora	24	Corpus Multilingual Children at Pre-school Age (MEKI)	unknown	yes	unknown	unknown	unknown	unknown	unknown	unknown	2004-2006	unknown	unknown	https://agd.ids-mannheim.de/MEKI_extern.shtml	registered for access -> got denied
Google search: speech children corpora	25	ALCEBLA	no	yes	unknown	unknown	23	9	14	unknown	2011	unknown	unknown	https://www.fdr.uni-hamburg.de/record/1530#.ZBQ71B-ZNhE	just children voices no adult voices; request sent; request granted, study supervisors voice included not seperated, no metadata available
Google search: speech children corpora	26	Common Voice	yes	no	18	unknown	unknown	unknown	unknown	unknown	2023	unknown	unknown	https://commonvoice.mozilla.org/en/about	downloaded Common Voice Corpus 13.0 from 15.03.2023, English language
Google search: speech children corpora	27	aGender	yes	yes	7	80	945	unknown	unknown	unknown	unknown	2010	German	https://aclanthology.org/L10-1178/	Martin.Eckert@telekom.de Wiebke.Johannsen@telekom.de 455,00€ fee https://www.phonetik.uni-muenchen.de/forschung/Bas/BasaGendereng.html voice quality was low due to recording through telephone

downloaded and most ideal corpus
downloaded for testing
considerable
requested and pending



NTNU

Norwegian University of
Science and Technology