Age Group Classification through Speech Recognition

1st Lippold, Marian
Student at Hochschule RheinMain
Hochschule RheinMain
Wiesbaden, Germany
marian.lippold@student.hs-rm.de

***Abstract*—In this paper, two transformer language models are analysed for their ability to assign spoken words to a specific age group. For this task, a dataset must be prepared that contains not only the transcribed words, but also information on the age and gender of the speakers. It is analysed how the models perform without fine-tuning and, if necessary, fine-tuning is carried out.**

***Index Terms*—Speech Recognition, Transformer, Artificial Intelligence**

Introduction

In today's world, where great advances in AI are being made through the introduction of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and the Attention Mechanism, the world is experiencing a new AI boom. Available datasets, whether validated and labelled or unvalidated, have become more and more extensive over time and enabling the training of AI models on a new scale. Voice-based offers and services, such as personal assistants in the domestics or business sectors or transcription and translation automation are used by most consumers and providers worldwide. This paper analyses and compares two popular language models in terms of their ability to determine the age behind the speaker.

Transformers

The architecture of the transformer models differs from the architecture of their predecessors, the Recurrent Neural Networks (RNN), in three important respects:

- Positional Encodings
- Attention
- Self-Attention

Positional Encodings

In the architecture of the RNNs, all words in a sentence were processed sequentially. The information about the word positions in the sentence was therefore ignored. Thanks to the innovation of positional encodings, the position is added to the input as a number when it is read in. The subsequent neural network therefore has information about the order of the words and can thus learn the relevance of word positions in the sentence structure.

Attention

There are parts of a sentence that are given a grammatical gender in a translation that was not present in the original representation. The attention mechanism allows the model to search for the parts of the sentence on which the grammatical gender depends in the original sentence and to carry out the translation depending on the influencing words. The model learns the information about which words are grammatically dependent on others during pre-training when it is trained with large amounts of data.

Self-Attention

The attention mechanism is applied here to the input in order to gain an understanding of the meaning of the words. For example, the word server in English can mean the network computer or the waiter in the restaurant. Self-attention means that the meaning of the word server in this context is already evaluated during the learning process based on the predicate and the object. This makes it possible to overcome word ambiguities, recognise idioms or determine the grammatical tense.

ASR-Transformer

The typical Automated Speech Recognition (ASR)-Transformer-model "maps an input-sequence, [...] the logMel filterbank feature, to a sequence of intermediate representations by the encoder." (Li et al., 2019) [23]. The Transformer architecture typically consists of Encoder and Decoder blocks, but can also be made up only of the Encoder. The Encoder processes the input sequences and generates representations from them. It consists of a stack of layers which are made of two main components:

- Multi-head self-attentation to focus on different parts of one input sequence when computing a representation to capture context.
- Feed-Forward Network to decrease linearity and emphasize important features while suppressing less relevant ones.

Backpropagation and Skip-Connections

In backpropagation, the parameters of a neural network are iteratively adjusted so that the loss function is minimised. Partial derivatives are used to determine how parameter changes

affect the loss function. Skip connections enable the output of one layer to be passed on not only to the next, but also to subsequent layers. This prevents the gradients, which are calculated multiplicatively, from becoming too small or falling to 0, which would stop the adjustment of the layers. [3]

Product Quantization

Product Quantization is a technique to approximate large datasets by decomposing the original high-dimensional space into a Cartesian product of lower-level dimensional subspaces. These are then quantized independently resulting in compact representations with reduced memory consumption. [19]

Beam Decoder

The beam decoder is a search algorithm that is used in sequence-to-sequence models. It is an approximate decoding method that examines several possible output sequences simultaneously and exists as a middle ground between greedy and exhaustive search [10].

Mixture of Experts

Jacobs, et al. [18] coin the term gating network, which selects from various local experts the one who can make the best prediction for the current input. For this purpose, experts are trained in the different regions of an input space and the space is completely covered. The gating network assigns weights to the experts, on the basis of which the predictions are selected.

Metrics

Accuracy: The accuracy metric measures the proportion of correctly classified instances out of the total instances, providing a simple yet effective way to evaluate a model's performance. Mathematically, it is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives that are correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score: The F1 score is a harmonic mean of precision and recall, providing a single metric that balances both false positives and false negatives. It is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

I. Related Work

The task of age and gender prediction from speech signals has been extensively studied over the past few decades. Various machine learning models have been developed and evaluated on different datasets, employing both traditional and deep learning techniques.

Data from Voices

Woods finds in her 1992 paper "It's not what she says, it's the way that she says it: the influence of speaker-sex on pitch and intonational patterns" [39] several differences between the sexes. Female and male voices differ significantly from each other. On one hand, female voices have higher amplitudes in the range from 0.96 to 3.36kHz, on other hand, their voices retain energy better at higher frequencies due to the lower overall spectral tilt.

When determining the age of a speaker, there are several factors that influence the waveform of speech. A distinction is made between prosodic characteristics, i.e. patterns in the accentuation of syllables and words, as well as rhythm, pauses and speed of speech. These characteristics are influenced by physiological factors such as the length and elasticity of the vocal cords, breathing control and lung volume, as well as the control of articulatory movements. [14] state that people of advanced age tend to speak more slowly. Men have more shimmer in their voices and tend to speak more softly than younger men. The frequency F0, also called (fundamental frequency), is the lowest frequency of a periodic waveform in the air transmission and is often related to the pitch of a voice. It increases with age in men, making their voice-pitch sound higher [32]. Women, on the other hand, would speak at a lower frequency in F0, i.e. generally lower, with less variation in this frequency and the ability to sustain certain tones for longer than younger women.

From a spectral perspective, i.e. the frequency distribution and energy of a speech signal, the noise-to-harmonics ratio shifts over the years, making a voice appear rougher. Age-related neuromuscular changes lead to reduced accuracy in speech production.

In general, Taylor, et al, [36] state that speech disorders occur more frequently and more clearly with age. (S. 1) The voice becomes more susceptible to phonetic instability, voice breaks and roughness with age. Taylor, et al., also state that the slower speech rate together with over-emphasis is used to compensate for decreasing precision in speech.

According to Woods, women tend to use stresses that first fall and then rise again. In general, they would use more rising and falling stresses, while men tend to emphasise monotonously and evenly. Women usually utilise the range of their voice pitch more than men.

Traditional Approaches

Early works on age and gender classification primarily relied on traditional machine learning techniques. Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) were popular choices, as seen in the studies using the aGender dataset during the 2010 Interspeech Paralinguistic Challenge. Lingenfelser et al. [25] reported an accuracy of 42.4% UAR on four age groups, and Katerenchuk [20] improved child speech detection by fusing acoustic and metadata features using similar models. Other traditional approaches, such as

those based on Linear Discriminative Analysis (LDA) and Hidden Markov Models (HMM), achieved moderate success in this domain [27].

Deep Learning Models

With the advent of deep learning, CNN and LSTM networks have shown promise in age and gender prediction tasks. Sánchez-Hevia et al. [35] and Tursunov et al. [37] applied CNNs to the CommonVoice dataset, achieving recalls of 76% and 74% UAR on different age-gender groups, respectively. Zazo et al. [40] utilized LSTM networks, reporting a Mean Absolute Error (MAE) of 6.58 years on the NIST test set.

Transformer Models

More recently, transformer-based models such as Wav2Vec and Whisper have gained popularity for this task. Gupta et al. [15] classified age using a wav2vec 2.0 model on the Timit dataset, achieving MAE values of 5.54 and 6.49 years for male and female speakers, respectively, using a the mixture of experts approach for male and female inputs. The study by Kwasny and Hemmerling [21] leveraged a QuartzNet architecture, which is an NVIDIA-developed CNN architecture, pre-trained on CommonVoice and VoxCeleb2, and fine-tuned for joint age and gender prediction, reaching performance levels comparable to other state-of-the-art models. Burkhardt, et al., (2023) use Wav2Vec2 to predict both age and gender simultaneously and achieve a mean average error of 8.35 years for all samples ages [9].

Challenges and Limitations

Despite the advancements, several challenges persist. The lack of standardized datasets and the variability in training-development-test splits across studies make direct comparisons difficult. Additionally, while deep learning models generally outperform traditional methods, their performance can degrade significantly when evaluated on out-of-domain data, as demonstrated by the cross-corpus evaluation results detailed in [9].

II. Models

A. Whisper

The OpenAI-developed Speech Recognition model "Whisper" was released in September 2022 and updated to V2 in December 2022 alongside a published paper by Radford et al. [30] The architecture of Whisper consists of the more common components of a transformer model: It takes input sequences sampled at 16,000Hz and computes a log-Mel spectogram on 25ms windows with a stride of 10ms. (S. 3) A spectogram is a visualisation of a Fourier-transformed signal over multiple windowed segments meaning the shift of loudness of the frequencies over time. The mel-scaled spectogram was developed by Volkmann, et al. in 1937 [38] to better depict the difference in pitch for a listener. The Convolutional layers and the GELU activation which process the log-mel-spectogram and extracts local features. (S. 3) Then sinusoidal position embeddings are added to preserve the order of input sequences. (S. 3) The Encoder of Whisper uses attention mechanisms and feed-forward-layers to create representations from the input sequences, while the Decoder uses cross-attention to generate the outputs. Cross-Attention means the splitting of attention on different parts of the input to generate an output. For the classification problem of this paper, the variant "small" with 244 million parameters is chosen due to ressource restrictions.

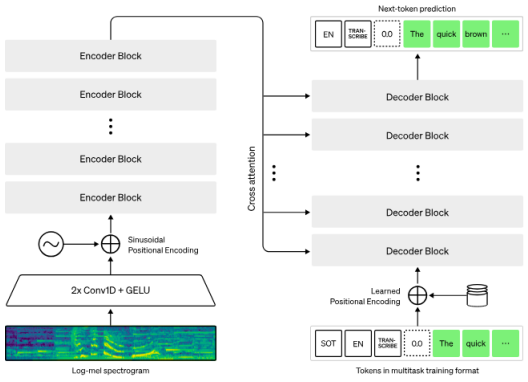

Fig. 1. Figure 1: Architecture of Whisper [28]

Wav2Vec

Wav2Vec is a language model developed by S. Schneider, et al. [34] developed at Facebook AI in 2019, which at that time was able to outperform the model with the best performance with lower data requirements and made promising progress in self-supervised training. (S. 1) It is a convolutional neural network, the convolutional architecture allows parallelisation of the model during pre-training. the model learns contextualised representations of the input (S. 1). The five-layer encoder network provides 30ms low-frequency feature representations as a result of 10ms audio. The subsequent context network combines 7 of the 30ms long latent representation into a single, contextualised tensor (S. 2). The tensor thus includes information about the moments before the sound and therefore also the surrounding sounds around the specific sound (S. 2). Skip connections are used within the context network. (S. 2) The top layer of the context network spans approximately 810ms, allowing the model to contextualise a large portion of the audio sequence (S. 2). Instead of the usual Mel filter banks, these representations are now passed to the acoustic models, where phonemes and words are generated from the representations (S. 3). For this purpose, various acoustic models such as 4-gram KenLM, [16], the word-based convolutional language model by Collobert, et al, [10] and the character-based convolutional language model by Likhomanenko, et al. [24] (p. 3) are used. The decoder architecture uses a contrastive loss function, which improves the model's ability to distinguish real audio signals from noise signals (S. 4). The model performs with a word-error rate of 2.43% on the Wall Street Journal dataset (S. 7). The authors compare the results of the models when using Mel filter banks with the results when using the representations learnt through pre-training. They find

an improvement in LER and WER when the representation practice introduced here is used to allow the model to learn representations of sounds and contexts.

B. Wav2Vec2

In 2020, Baevski, et al. further develop the model with Wav2Vec2 [6] by passing the output from the feature encoder into a context layer based on the Transformer architecture. Unlike the traditional way of determining fixed positions of the embeddings, a convolutional layer is used to encode the relative positions of the embeddings and pass them to the Transformer layer (S.2). The context network here is the transformer layer, which is used to find dependencies across the entire sequence using the self-attention mechanism via continuous language representations (S.2). This version of the model also has a quantisation module, which converts the output of the feature encoder into a finite set of language representations using product quantisation (S.3). The Gumbel-Softmax algorithm enables the selection of discrete codebook entries that contain the representations. These can sometimes represent the same sound in several languages (S.3). The representations are learnt using a contrastive loss function and the diversity loss encourages the model to use the different codebook entries equally often (S.3). This model is also first pre-trained with a larger dataset of unlabelled data and then fine-tuned with labelled data. The model achieves a word-error rate of 1.8/3.3 on the full LibriSpeech dataset.

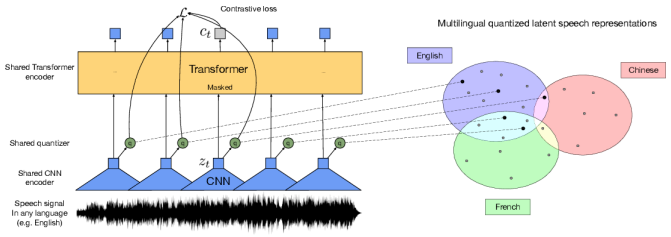

Fig. 2. Figure 2: Simplified representation of the encoder-architecutre of Wav2Vec2 [11]

XLSR-Wav2Vec2: Wav2Vec2-XLSR is a model developed by Facebook AI based on Wav2Vec2 by A. Conneau, et al. [6], which was trained using 53 different languages. The transformer encoder learns the representations of loud on all languages as it is shared between all languages. The quantisation module is also shared between all languages.
Wav2Vec2-large-robust-6-ft-age-gender: A Wav2Vec2 model fine-tuned to estimate age and gender from Burkhardt, et al. [9]. For simplicity, this model will be called "Wav2Vec2 Age-Gender-Robust" from now on.

III. Datasets

This chapter details the datasets which were discovered during research. The data for this paper must be available in German and contain information on the age and gender of the speakers. As the only dataset which has been used in training and evaluation was Mozilla CommonVoice, this chapter also details why other datasets could not be used.

CommonVoice-Mozilla

CommonVoice [4] is a dataset that is spoken and confirmed by voluntary contributors. The project is hosted by Mozilla. According to its own information, the project aims to combat the fact that English-speaking white men make up the majority of speakers in order to improve the quality of speech recognition for all people. The version of the dataset used for this paper, 18.0, comprises 1431 hours of speech material from 19146 speakers. The age distributions can be read from Figures 3, 4 and 5.

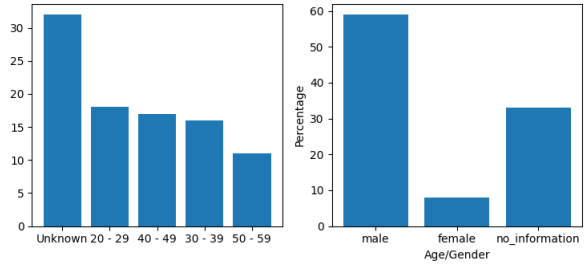

Fig. 3. Figure 3: Age range and distribution of CommonVoice 18.0 according to the publishers

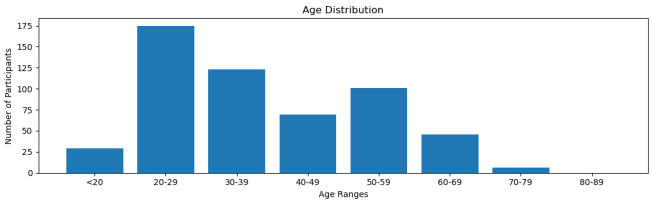

Fig. 4. Figure 4: Age range of female participants as analysed from the dataset

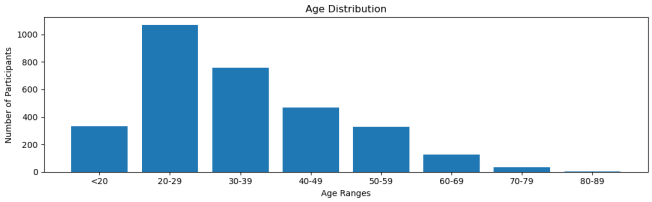

Fig. 5. Figure 5: Age range of male participants as analysed from the dataset

A. German Distant Speech Data Corpus

The dataset GermanSpeechDat [29] was recorded by the University of Hamburg in cooperation with the Dialog+ project. It contains 36 hours of speech material from 180 speakers, 130 of whom are male and 50 female. The age distribution is approximately 20 in 18-20, 20 in 31-40 and 100 in 21-30. Unfortunately, it is not clear from the available data how the audio files can be assigned to the metadata [22].

B. (German) SpeechDat II

The SpeechDat II Corpus [33] was collected by the BAS of Munich University through subcontracts with Siemens and

Vocalis for landline and mobile network telephony. This is available for academic purposes at ELRA. It contains 3 age groups, 16-30, 31-45, and ¿45, with a gender distribution of about 1:1 and contains 5000 speakers. The sampling rate is 8000Hz and was recorded in the office, at home or in a telephone booth. Unfortunately, the assignment of file and age of the person is no longer available in the dataset, so that a prediction can never be tested [17].

C. LibriVox

This source is freely accessible and offers the option of selecting the language of the audio files. However, the source fails in the closer selection due to the lack of age information. In the course of this study, an attempt was made to personally ask German-speaking members about their age but only a single person out of around 25 respondents replied. It would be conceivable to contact sources with a public email address or public presence, such as their own website, separately in order to recruit them for a dataset.

D. EmoDB

The EmoDB database [8] contains transcribed data categorised by emotion from 10 actors act in different emotional states while speaking. The age range of the dataset is 21-35 years, which is already overrepresented. However, since the websites that offer [5] and [1] for download do not load correctly at the time of writing, they cannot be used.

E. Bundestag Archive

More potential data sources are reports from public media that offered and the Bundestag archive, which has archived the plenary sessions completely as video since 1995. Since the people on these recordings are well-documented public figures, it should be possible to extract the age and gender features for the speaking persons. Also, this data will likely be noisy, as the speeches can get interrupted by verbal expressions of the parliamentarians present.

F. aGender

Dataset published by Burkhardt, et al. (Burkhardt, et al., 2010) [7] with 47h German language of 954 participants stored in 8kHz. Due to ELRA restrictions, the author will not use this data. Is currently not available due to problems with the website [2] and requires an academic licence, which the author has not received for this paper.

G. Datenbank für Gesprochenes Deutsch

This maintainer has 41 datasets of spoken German words, but these are not licensed for the training of person-recognising AI models or for reuploading to other portals or Google Drive.

IV. Conception and Implementation

The following chapter details the Conception and Implementation of the study. First, the steps taken to sight and pre-process the Mozilla CommonVoice dataset are described. Then, the individual process of training is described for each model used.

A. Pre-Processing

After collecting and viewing the data records, the CommonVoice dataset is first analysed and cleansed due to a lack of premises regarding data quality. To do this, all invalid values in the Age, Gender and Accent columns are removed along with the row. All duplicated entries are then removed using the client_id column in order to obtain a graphical overview of the distribution of people by gender and age group. The age groups correspond to the decades from ten to 89 years of age. The clips of all speakers are read out again and saved in order to identify overlaps between the train.tsv dataset and the validated.tsv, test.tsv and dev.tsv datasets. As significant overlaps were detected in validated.tsv, the datasets train.tsv and dev.tsv are merged. The dataset is now available as a path to the audio file and metadata for the file and is first made into a dataset object with predefined features (genders, age groups and all accents) using the huggingface-datasets library. This object is then saved as an .arrow dataset and uploaded to Huggingface to enable retrieval from Google Colab. Resampling of the data is not necessary as it is already available at a sampling rate of 16,000Hz.

A dataset with a single sample from each participant is extracted and described in Table 1. Another dataset with a balanced amount of clips in each age-class is extracted and described in Table 2. Both datasets are uploaded to Huggingface [26].

Age Group	Participants
Teens	363
Twenties	1246
Thirties	879
Fourties	538
Fifties	427
Sixties	174
Seventies	42
Eighties	3

TABLE I

TABLE 1: PARTICIPANT NUMBERS IN UNBALANCED DATASET

Age Group	Clips
Teens	492
Twenties	499
Thirties	500
Fourties	499
Fifties	495
Sixties	484
Seventies	422
Eighties	26

TABLE II

TABLE 2: AUDIO CLIP NUMBERS IN BALANCED DATASET

B. Training - Wav2Vec2-XLSR

The necessary libraries of Huggingface, ”datasets” and ”transformers” for working with the data and models, as well as torch for performing deep learning operations and torchaudio for processing speech data are installed. Furthermore, ”numpy” and ”sklearn” are installed to calculate metrics, as a documented bug occurs when calculating the accuracy [12].

The dataset is prepared for training by the processor extracting the features from the respective audio signal in a ”map”-function and assigning the respective age label. The variable length of the samples is then handled in the DataCollator by padding them to a uniform length. If the ”input_values” and ”labels” are not yet available as tensors, they are converted into the form of tensors by the DataCollator at the latest.

The calculation of the metrics is defined. The metrics selected are Accuracy, to determine the correctly determined samples, the macro-averaged F1-Score, which calculates the F1-Score individually for each age group, the Classification Report, which calculates and clearly displays general statements about the model performance, and the Confusion Matrix, which determines the true and false positives and displays them in tabular form. Furthermore, the metrics mean-squared-error and mean-absolute-error are calculated so that the distance between the predicted and actual age class is shown numerically. The hyperparameters are not modified for the time being and correspond to the standardised values for this type of model and task, which are intended to prevent overfitting of the model. Only the ”batch_size” parameter is set to 9 in order to optimally utilise the resources from the working environment.

A modified Wav2Vec2 model with a classification header is used to perform the task, which was developed by Reiser, L. & Fivian, R. published in 2021 [31]. The classification header consists of two fully meshed layers of 1024-dimensions and behind it another layer which maps the 1024-dimensional data to the 8 age classes. The reasoning for using two layers instead of a singular one comes from the bachelor thesis of Reiser & Fivian [31], who identify the variability of the shape of the 3D-output vector from wav2vec2. The dimension of ”output_features” varies based on the length of the input signal. The authors take the mean from it to receive a fixed-size tensor of [batch_size, 1024]. By adding an Hyperbola-tangent activation function, the authors introduce non-linearity and with the second classification layer, the output is reduced to the number of classes.

For the extraction of features, the mean of the hidden states is calculated over the sequence length, which pools the features across time. These fixed-size feature vectors are passed to the classification layer.

For a loss-function we use Cross-Entropy-Loss as it is standard for multi-class classification tasks [13]. The FeatureExtractor of Wav2Vec2 is frozen to prevent updating the feature extraction layers during training. This is done because we only want to train the classification layers and not the rest of the model for comparison.

C. Training - Wav2Vec2-Age-Gender-Robust

The training of this model follows the same pipeline as the training of the XLSR-variant, yet utilises the Wwav2vec2-Age-Gender-Robust-model.

D. Training - Whisper

The training pipeline of Whisper is essentially identical to the pipeline written for the Wav2Vec2 models. Whisper uses a sampling-rate of 16000Hz just like Wav2Vec2. However, the pipeline is written without explicit usage of an Processor-object. Instead, the audio-data is directly fed to the FeatureExtractor which returns the correct data-format for the training of Whisper. Training is done in 10 epochs and with a learning rate of 5e-5 because the model is already pre-trained. Cross-Entropy-loss is used as the loss-function.

V. Performance and Evaluation

For evaluation, the ”evaluate”-method of Huggingface's Trainer class is used, but the scores of the training steps are being considered as well. The dataset used for evaluation consists of 10% of the respectively used subset which the model has not seen during the training and validation cycle.

A. Evaluation - Wav2Vec2-XLSR

The first conducted test which used a dataset with unbalanced classes showed no clear signs of the model getting significantly better at classifying. The training loss as well as the validation loss stay almost constant, while the F1-Score increases from 7% to 15% and the mean errors decrease slightly. It is worth noting the Wav2Vec2-XLSR-variant trained significantly slower than the Wav2Vec2-Age-Gender-Robust, not training the full dataset with the available amount of gpu time.

Step	Training Loss	Validation Loss	Accuracy	F1 Score	MAE^M
100	1.666300	1.702984	0.340599	0.072590	2.285714
600	1.775500	1.681883	0.354223	0.120505	2.021948
1600	1.639200	1.663493	0.340599	0.072590	2.285714
2500	1.542500	1.633940	0.356948	0.097525	2.085996
3000	1.645700	1.605005	0.359673	0.153412	1.724660

TABLE III
TABLE 3: SUMMARY OF TRAINING AND VALIDATION METRICS AT SELECTED STEPS FOR WAV2VEC2 ON A UNBALANCED SET

The second conducted test on a balanced dataset had a runtime of around 285 minutes. This was the only training run that would run out of memory, so it was repeated, leading to lesser accuracy and worse overall statistical results, even though the conditions were the same. The first run showed significant improvements in all categories, reaching an accuracy score of 68.22%. The second run showed signs of overfitting again and the accuracy stayed at around 60%.

Step	Training Loss	Validation Loss	Accuracy	MAE
100	1.968900	1.961521	0.169096	1.609329
600	1.412200	1.281453	0.510204	0.994169
900	0.880000	1.295183	0.591837	0.737609
1300	0.677400	0.945514	0.661808	0.489796
1500	0.479700	0.981552	0.682216	0.513120

TABLE IV
TABLE 4: SUMMARY OF TRAINING AND VALIDATION METRICS AT SELECTED STEPS FOR WAV2VEC2-XLSR ON A BALANCED SET

The evaluation results of Wav2Vec2-XLSR denoted in the following table. It is clearly visible that the classification task

only performed correctly when trained with a balanced dataset. While the accuracy and F1-Score are generally high, the model was challenged the most with correctly classifying the class ”twenties”.

Metric	XLSR-Balanced	XLSR-unbalanced
Teens F1-score	0.90	0.00
Twenties F1-score	0.84	0.33
Thirties F1-score	0.91	0.18
Fourties F1-score	0.98	0.28
Fifties F1-score	0.92	0.12
Sixties F1-score	0.96	0.00
Seventies F1-score	0.99	0.00
Eighties F1-score	1.00	0.00
Accuracy	0.93	0.20
F1 Macro Avg	0.94	0.11
Eval Loss	0.51	2.41
Eval MSE	0.22	4.80
Eval MAE	0.11	1.70
Eval MAEM̂	0.10	1.99

TABLE V

TABLE 5: SUMMARY OF EVALUATION METRICS FOR WAV2VEC2-XLSR TRAINED ON BALANCED AND UNBALANCED DATASETS

B. Evaluation - Wav2Vec2-Age-Gender-Robust

The first conducted test which used a dataset with unbalanced classes showed signs of the Wav2Vec2-Age-Gender-Robust model overfitting the data. While the training-loss decreases from 0.6949 in step 10 to 0.0035 in the final step 48, the validation loss increases from 1.8248 to 6.0960. This is a sign of overfitting the data. The accuracy in the beginning is around 39% and in the end is around 44%, however, the accuracy is varying between 39% and 47% during training and does not significantly improve overall.

Step	Training Loss	Validation Loss	Accuracy	F1 Score
100	1.653700	1.543556	0.395095	0.117598
600	1.279300	1.464648	0.405995	0.366613
1200	0.724200	2.168830	0.395095	0.259291
1800	0.532100	2.834368	0.457766	0.320133
2400	0.172900	3.880411	0.457766	0.317116

TABLE VI

TABLE 6: SUMMARY OF TRAINING AND VALIDATION VETRICS AT SELECTED STEPS FOR WAV2VEC2-AGE-GENDER-ROBUST ON A UNBALANCED SET

The second conducted test which used a dataset with balanced classes also showes signs of overfitting with the training loss decreasing and the validation loss increasing. However, the accuracy achieved at the last training step is significantly higher than at the first step with 0.760933% to 0.311953%. During training, the accuracy rises to 70% at the ninth step, doubling the accuracy in the first five steps.

The evaluation on unseen data yields results comparable to Wav2Vec2-XLSRs evaluation, showing a clear division between training on balanced and unbalanced data. Furthermore, the model achieves a minimum of 0.92 on the F1-Score on all age-classes with a overall accuracy of 94%.

C. Evaluation - Whisper

The training conducted on Whisper is only done with the balanced dataset, as training and evaluation results suggest that a unbalanced dataset is not fitting for a classification problem. While training on balanced data, Whisper scores comparable to the XLSR-variant trained on the unbalanced dataset, while requiring only 45 minutes to train.

Step	Training Loss	Validation Loss	Accuracy	MAE^M
100	1.707100	1.659937	0.311953	1.227939
200	1.280000	1.217010	0.510204	0.743291
600	0.665300	0.954688	0.676385	0.531457
1000	0.452000	1.024279	0.711370	0.381787
2000	0.112300	2.191491	0.717201	0.354784
4000	0.011300	2.521255	0.763848	0.278057

TABLE VII

TABLE 7: TRAINING AND VALIDATION METRICS AT SELECTED STEPS FOR WAV2VEC2-AGE-GENDER-ROBUST ON A BALANCED SET

Metric	Balanced	Unbalanced
Teens F1-score	0.92	0.13
Twenties F1-score	0.92	0.66
Thirties F1-score	0.92	0.43
Fourties F1-score	0.92	0.30
Fifties F1-score	0.94	0.33
Sixties F1-score	0.95	0.00
Seventies F1-score	1.00	0.00
Eighties F1-score	1.00	N/A
Accuracy	0.94	0.47
F1 Macro Avg	0.95	0.23
Eval Loss	0.65	2.55
Eval MSE	0.19	1.35
Eval MAE	0.10	0.75
Eval MAEM̂	0.09	1.28

TABLE VIII

TABLE 8: SUMMARY OF EVALUATION METRICS FOR WAV2VEC2-AGE-GENDER-ROBUST TRAINED ON BALANCED AND UNBALANCED DATASETS

Epoch	Training Loss	Validation Loss	Accuracy	MAE
1	1.511400	1.420326	0.440233	0.927114
3	0.665200	1.153432	0.626822	0.588921
5	1.341400	1.999827	0.641399	0.632653
9	0.000000	2.973631	0.685131	0.507289
10	0.000000	3.073634	0.688047	0.501458

TABLE IX

TABLE 9: SUMMARY OF TRAINING AND VALIDATION METRICS FOR SELECTED EPOCHS OF WHISPER

Evaluation shows that Whisper performs significantly worse than the Wav2Vec2-variants on the balanced dataset. Whisper is not reliable in classifying the age-classes below the age of sixty, at and above sixty however, there is a F1-Score of at least 0.87 for every class. Notably, the loss-function on the evaluation dataset reports a loss five to six times higher than in the Wav2Vec2-variants.

D. Discussion

Above all, the results show that a balanced dataset is essential for training a classifier. In a dataset whose audio files were randomly selected from the existing ones of an age group, all three models classify the older age groups quite reliably. This suggests that the voices of older people in particular can be reliably distinguished from the voices of younger people. This result is supported by the basic research of Goy, et al. [14], Woods [39] and Taylor, et al. [36].

Metric	XLSR-Balanced	Age-Gender-Robust	Whisper
Teens F1-score	0.90	0.92	0.61
Twenties F1-score	0.84	0.92	0.56
Thirties F1-score	0.91	0.92	0.52
Fourties F1-score	0.98	0.92	0.74
Fifties F1-score	0.92	0.94	0.67
Sixties F1-score	0.96	0.95	0.87
Seventies F1-score	0.99	1.00	0.90
Eighties F1-score	1.00	1.00	1.00
Accuracy	0.93	0.94	0.69
F1 Macro Avg	0.94	0.95	0.73
Eval Loss	0.51	0.65	3.10
Eval MSE	0.22	0.19	1.26
Eval MAE	0.11	0.10	0.53
Eval MAEM̂	0.10	0.09	0.45

TABLE X
TABLE 10: COMPARISON OF XLSR-BALANCED, NEW EVALUATION, AND WHISPER EVALUATION RESULTS

While the results for Whisper with an average accuracy of 69% do not correspond to the hoped-for values, the Wav2Vec2 variants classify quite reliably on the dataset. However, it must be said that no cross-corpus evaluation was carried out to check these results on completely new datasets. This means that overfitting on the existing data cannot be ruled out.

In general, it must be noted that in the classification into a total of 8 age groups, all models identify correctly more often than they are wrong. An adaptation of the task or a simplification of the age classes into ”young”, ”adult” and ”old”, on the other hand, should enable more reliable classification due to the striking vocal differences.

The differences in the architecture of Whisper and Wav2Vec2 help to explain the results. Whisper is designed for speech recognition and translation tasks, whereas Wav2Vec2 has already been used successfully for classification tasks. This confirms that Whisper is therefore not suitable for reliable classification. On the other hand, Whisper completed the training about two and a half times faster than Wav2Vec2-Age-Gender-Robust, taking about 45 minutes. If only limited resources are available and the precision of the classification does not have to be completely reliable, it would be worth considering using Whisper.

As there is almost no data from 80+ year olds, who will certainly be represented in the general public, e.g. nursing homes, problems could arise in practice in nursing homes. In addition, most of the data are read-in sentences with little background noise. The distribution between male and female is also problematic, so that either a lot of data is lost during balancing data is lost or the male speakers are overrepresented. The data is recorded in a neutral tone and attention was paid to intelligibility, which will not be the case in practice. This makes transferring the results to the real world challenging.

Also, even though the dataset was balanced as best as possible into the age-classes, there are only 26 clips or 3 participants in the age-class of ”eighties”, which means it is significantly underrepresented. Furthermore, there was not balancing between the genders, nor was the dataset split into a male and female subdataset. Doing this might further improve the results.

E. Future Work

The idea of this paper was to compare two popular Speech-Recognition models without any further fine-tuning on their ability to classify speech in age-classes. It is aimed to determine the more suitable architecture from these models which have already been pre-trained unsupervised. Also this paper showed the multitude of available datasets which in the future could be leveraged by other researchers.

Further experimentation may be done by handling the problem as a regression problem instead of a classification problem which is not in the scope of this paper. For the Wav2Vec2 models, experimentation with other activation functions and classification headers might yield better results. Also, further experimentation with the hyperparameters of training might be conducted to optimise the current results further.

Another change to consider is to train the model on gender-specific data and to employ a mixture-of-experts mechanism in a final model which evaluates in the real world.

Conclusion

In the work done in this paper, the author researched multiple open-source, non-commercial, age- and gender-labeled, german datasets. Research uncovered multiple potential data sources and with respect to the available ressources, the most fitting, Mozilla CommonVoice was chosen. The dataset was examined, pre-processed and divided into sub-datasets which fit the available computation times. Age and Gender distributions were computed which deviated from the information on the official website. The Speech Recognition models ”Wav2Vec2-XLSR” and ”Wav2vec2-large-robust-6-ft-age-gender” based on the Wav2Vec2-architecture were extended with an classification head and trained. The Speech Recognition model ”Whisper” was trained as a classifier. Evaluating the models, the Wav2Vec2-architecture performs more accurate than the Whisper-architecture, while the Whisper model took significantly less training time. No cross-corpus testing was conducted and overfitting on the dataset could not be ruled out.

All models and datasets the author created for this paper are publicly accessible on Huggingface [26].

REFERENCES

[1] Berlin database of emotional speech (emo-db). Accessed: 2024-08-17.

[2] Agender: Gender and age classification dataset, 2013. Accessed: 2024-08-17.

[3] N. Adaloglou. Intuitive explanation of skip connections in deep learning. *https://theaisummer.com/*, 2020.

[4] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus, 2020.

[5] audEERING GmbH. Berlin database of emotional speech (emo-db). Accessed: 2024-08-17.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[7] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. A database of age and gender annotated telephone speech. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. volume 5, pages 1517–1520, 09 2005.

[9] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller. Speech-based age and gender prediction with transformers, 2023.

[10] R. Collobert, A. Hannun, and G. Synnaeve. A fully differentiable beam search decoder, 2019.

[11] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979, 2020.

[12] H. Face. Issue: Pre-trained model questions. https://github.com/huggingface/setfit/issues/228, 2023. Accessed: 2023-09-13.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[14] H. Goy, M. Pichora-Fuller, and P. Van Lieshout. Effects of age on speech and voice quality ratings. *The Journal of the Acoustical Society of America*, 139:1648–1659, 04 2016.

[15] T. Gupta, D.-T. Truong, T. Anh, and C. Siong. Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model, 03 2022.

[16] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn. Scalable modified Kneser-Ney language model estimation. In H. Schuetze, P. Fung, and M. Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics.

[17] L.-M.-U. M. Institut für Phonetik und Sprachverarbeitung. The bits speech synthesis cookbook. Accessed: 2024-08-17.

[18] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[19] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33:117–28, 01 2011.

[20] D. Katerenchuk. Age group classification with speech and metadata multimodality fusion. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 188–193, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

[21] D. Kwasny and D. Hemmerling. Joint gender and age estimation based on speech signals using x-vectors and transfer learning, 2020.

[22] U. H. Language Technology Group. Acoustic models for speech recognition. Accessed: 2024-08-17.

[23] S. Li, D. Raj, X. Lu, P. Shen, T. Kawahara, and H. Kawai. Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation. In *Interspeech*, 2019.

[24] T. Likhomanenko, G. Synnaeve, and R. Collobert. Who needs words? lexicon-free speech recognition. In *Interspeech 2019*, interspeech 2019. ISCA, Sept. 2019.

[25] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. Andre. Age and gender classification from speech using decision level fusion and ensemble based techniques. pages 2798–2801, 09 2010.

[26] M. Marian. Hugging face profile - mlmarian. https://huggingface.co/mlmarian, 2024. Accessed: 2024-09-20.

[27] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. Bauer, and B. Littel. Comparison of four approaches to age and gender recognition for telephone applications. volume 4, pages IV–1089, 05 2007.

[28] OpenAI. Introducing whisper. https://openai.com/index/whisper/, 2022. Accessed: 2024-09-11.

[29] S. Radeck-Arneth, B. Milde, A. Lange, E. Gouvea, S. Radomski, M. Mühlhäuser, and C. Biemann. Open Source German Distant Speech Recognition: Corpus and Acoustic Model. In *Proceedings Text, Speech and Dialogue (TSD)*, pages 480–488, Pilsen, Czech Republic, 2015.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[31] L. Reiser and R. Fivian. Speech classification using pre-trained transformer models. https://www.zhaw.ch/storage/engineering/institute-zentren/cai/BA21_Speech_Classification_Reiser_Fivian.pdf, 2021. Bachelor's thesis.

[32] ResearchGate Community. Who knows the relationship among pitch, fundamental frequency (f0), and tone/intonation of voice? https://www.researchgate.net/post/Who-knows-the-relationship-among-pitch-fundamental-frequency-F0-and-tone-intonat 2015. Accessed: 2024-09-13.

[33] F. Schiel, C. Draxler, A. Baumann, T. Ellbogen, and A. Steffen. *The Production of Speech Corpora*. 2012.

[34] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019.

[35] H. Sánchez-Hevia, R. Gil-Pita, M. Utrilla, and M. Rosa-Zurera. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 81, 01 2022.

[36] S. Taylor, C. Dromey, S. L. Nissen, K. Tanner, D. Eggett, and K. Corbin-Lewis. Age-related changes in speech and voice: Spectral and cepstral measures. *Journal of Speech, Language, and Hearing Research: JSLHR*, 63(3):647–660, 2020.

[37] 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*, 21(17), 2021.

[38] J. Volkmann, S. S. Stevens, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3_Supplement):208, 1937. Accessed: 2024-09-13.

[39] N. Woods. It's not what she says, it's the way that she says it: the influence of speaker-sex on pitch and intonational patterns. *Journal of Sociolinguistics*, 13(1):85–97, 1989.

[40] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak. Age estimation in short speech utterances based on lstm recurrent neural networks. *IEEE Access*, 6:22524–22530, 2018.