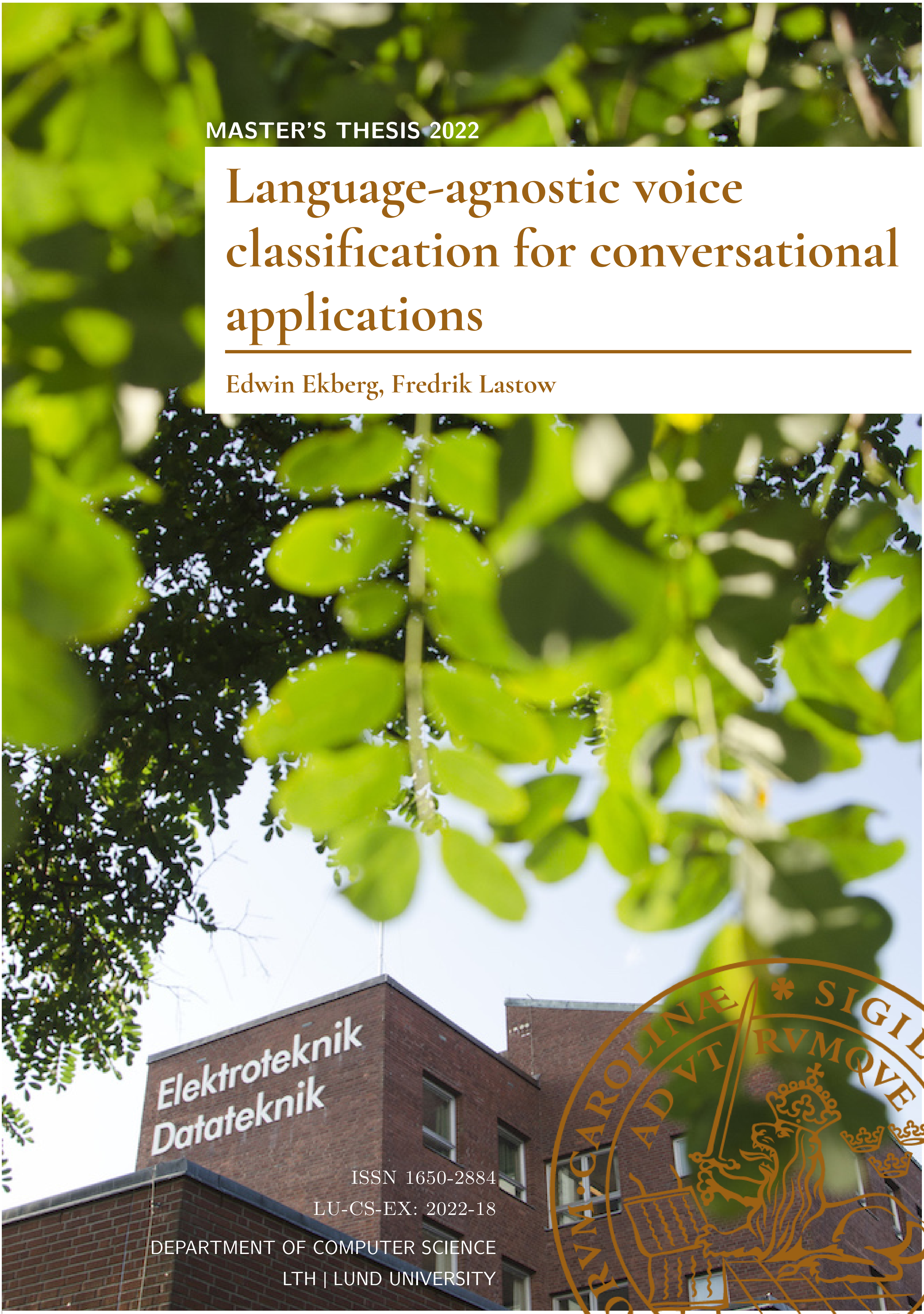
MASTER'S THESIS 2022

Language-agnostic voice classification for conversational applications

Edwin Ekberg, Fredrik Lastow



ISSN 1650-2884
LU-CS-EX: 2022-18

DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2022-18

Language-agnostic voice classification for conversational applications

Språkoberoende klassificering av röst för konversationsbaserade applikationer

Edwin Ekberg, Fredrik Lastow

Language-agnostic voice classification for conversational applications

Edwin Ekberg
edwin.ekberg@gmail.com

Fredrik Lastow
fredriklastow@gmail.com

April 6, 2022

Master's thesis work carried out at Sinch AB.

Supervisors: Michael Truong, michael.truong@sinch.com
Pierre Nugues, pierre.nugues@cs.lth.se

Examiner: Jacek Malec, jacek.malec@cs.lth.se

Abstract

For audio-based conversational applications, adapting responses to the attributes of the correspondent is an integral part in making the conversations sound natural. Two speaker attributes that humans can estimate quite well, based solely on hearing a person speak, is the gender and age of that person.

In the field of speech processing, age and gender classification are relatively unexplored tasks, especially in a multilingual setting. In most cases, hand-crafted features, such as MFCCs, have been used with some success. However, recently large transformer networks, utilizing self-supervised pre-training, has shown promise in creating general speech embeddings for various speech processing tasks.

We present a baseline for gender and age detection, in both monolingual and multilingual settings, for multiple state-of-the-art speech processing models, fine-tuned for age classification.

We created four different datasets with data extracted from the Common-Voice project to compare monolingual and multilingual performances. For gender classification, we could reach a macro average F1 score of ~96% in both a monolingual and multilingual setting. For age classification, using classes with a size of 10 years, we obtained a macro average mean absolute class error (MACE) of 0.68 and 0.86 on monolingual and multilingual datasets, respectively.

For the English TIMIT dataset, we improve on the previous state of the art for both age regression and gender classification. Our fine-tuned WavLM model reaches a mean absolute error (MAE) of 4.11 years for males and 4.44 for females in age estimation and our fine-tuned UniSpeech-SAT model reaches a macro average F1-score of 0.998 for gender classification.

In order to improve the performance of the pre-trained state-of-the-art speech processing models we applied transfer learning. This increased performance for both gender and age classification. Thus, the networks successfully became better at distilling speaker related information out of voice clips through fine-tuning.

All the models were deemed fast enough on a GPU to be used in a real-time settings, and accurate enough to be applicable in multilingual conversational applications.

Keywords: Age estimation, gender classification, multilingual, speech processing, pre-trained, fine-tuning, embeddings

Acknowledgements

First and foremost, we would like to thank our supervisor Pierre Nugues, for giving us great guidance throughout this project. Pierre has been very patient and helpful through our many iterations of models and datasets. He has helped us with all aspects of our thesis, both with decisions in our experiments and practical work, but also significantly in writing and proof reading. Without Pierre, I think we would still be running our "final" experiments.

A big thank you is also due to our supervisors and friends at Sinch for giving us the opportunity to do this project with them. They've all given us great insights, which have been proven invaluable to our results. Thank you to both Michael Truong, Simon Åkesson and Paulo Fonseca for the support and help in what ever capacity we needed, you always came through. Thank you to Pieter Buteneers for listening to our progress every week and steering us in the right direction with your great advice.

We would also like to thank our examiner Jacek Malec and our opponents for reviewing our work and letting us prove its worth.

Contents

Chapter 1
Introduction

Artificial intelligence and machine learning are consistently replacing humans in tasks that are simple enough for a machine to deal with. However, these tasks are increasingly getting more complex due to improvements in the methods and algorithms within machine intelligence, as well as increasing computational resources. One such task, on the brink of being able to be automated, is conversation.

An important part when trying to make conversations with a machine sound natural, is to mimic the way humans would approach the conversations. A human typically has a good sense of whom they are speaking with and can observe the other party's personal attributes. This information is then used in order to adapt one's answers to the other party, making the conversation sound more natural.

Specifically, human beings instinctively know a lot about another person based solely on their voice. In the context of conversational applications, such as customer services, support and commerce, adapting to speaker attributes could play an important role in delivering a high quality user experience. By incorporating knowledge of the user into the voice bots, one enables them to adapt their responses accordingly. Thus, they can provide better service and an improved experience.

Two speaker attributes that humans can estimate quite well by only hearing a person speak, is the gender and the age of the person. Thus, we set out to build a machine learning model to classify a person's age and gender based solely on their voice. Ideally, this model should also be language-agnostic and make correct classifications regardless of the language spoken.

1.1 Previous work

Historically, hand-crafted features combined with classical machine-learning techniques, such as linear discriminant analysis, K-nearest neighbors, CART, random forests and support vector machines, have been used with some success in order to classify speaker attributes in voice. However, in recent years, advancements in artificial neural networks and deep learning (mainly due to increased computational capacity) has improved the state of the art in this domain (Kwasny and Hemmerling, 2021). Thus, the hand-crafted features are becoming less efficient compared to large embedding networks that are constructing more useful features by themselves. Below is a description of some of the most important studies throughout this journey, pertaining to gender and age detection in voice.

In the field of speech categorization, speaker identification is a more prevalent speech processing task than gender and age detection. In the next section, we therefore provide a description of the state of the art within speaker identification, given that this field is larger, with more activity and that the techniques are relatively applicable to both gender and age detection.

1.1.1 Speaker attribute classification

Shue and Iseli (2008) found that features related to the voice source, i.e. the physiological properties of the glottis and the vocal tract, can be used to improve the performance of gender detection systems. In particular, they successfully used voice source measures such as formant frequencies, formant bandwidths, open quotient and source spectral tilt correlates to increase gender classification accuracy.

In a survey paper, Mishra and Shukla (2017) conclude that pitch and formant frequencies are unsatisfactory features due to their sensitivity to noise. Instead different cepstral domain features, such as Mel frequency cepstral coefficients (MFCC) and RASTA-PLP have proven more successful in gender detection systems.

Qawaqneh et al. (2017) suggest that the most important feature set within gender and age classification has been the MFCCs, since they capture the part of the spectrum that is related to the vocal tract and filter out the prosodic information, i.e. what consonants are pronounced and when. Furthermore, the authors suggest that DNNs have great potential to improve these features, just like they have done in computer vision. Specifically, they suggest the I-vector as a potential technique, as it was the state-of-the-art technique within speaker recognition and language detection at the time of the article (in 2017).

Alkhawaldeh (2019) states that the most commonly utilized feature sets within voice gender recognition are mel-scaled power spectrogram (Mel), mel-frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma), spectral contrast (Contrast), and tonal centroid features (Tonnetz). He concludes that the MFCC, Chroma, and Mel features are the best for this task and yield similar results as they are related to each other. This author also writes that the most efficient classifiers and feature extractors of superior accuracy on voice gender recognition include deep neural networks (DNNs) and convolutional neural networks, which is consistent with the notion that these are becoming good enough to create their own features, rather than relying on, for example, MFCCs.

Nasef et al. (2021) outline the importance of feature extraction in voice gender recognition, and how spectral features, such as MFCC, have outperformed more classical pitch-based approaches. Their study concludes that the biggest shortcoming among these feature engineering techniques is their poor performance in noisy environments, which is expected to be better with more complex deeper networks using attention. The study introduces two networks on top of MFCC features that use self-attention and claims to achieve state-of-the-art performance in noisy environments for voice gender recognition, with an accuracy of 96% on

VoxCeleb.

Overall, the gender detection accuracy is quite high. The best results are ranging from 87% to 100%, depending on the dataset, in particular its noise level. For example, Nasef et al. (2021) reached 96% on the noisy VoxCeleb, whereas Kwasny and Hemmerling (2021) get 99.6% on the popular, yet much cleaner dataset, TIMIT.

Research on age detection is sparse, as most research rather focuses on speaker identification and verification. However, Kwasny and Hemmerling (2021) seem to have achieved the best published results on the TIMIT dataset. They achieve a 5.12 mean absolute error (MAE) for males and 5.29 MAE for females by training a QuartzNet embedder (using MFCC features) with a two-stage transfer learning scheme (pre-training on both CommonVoice and VoxCeleb). Importantly, the paper shows the usefulness of deep, residual, convolutional architectures for this type of task.

1.1.2 MFCCs and F-bank features

Since MFCCs are one of the most important feature sets within voice classification, this section provides a quick explanation of them. The conversion of an audio signal into MFCCs consists of the following steps:

1. **Pre-emphasis**. This step emphasizes higher frequencies in order to make the sound more like the true spectrum directly from the vocal tract.

2. **Windowing**. A Hanning or Hamming window of 20-25 ms is panned over the signal. Importantly the windows should be overlapping, ensuring all parts of the sound signal are close to the center of some window. The rest of the calculations is done for each frame respectively.

3. **Discrete Fourier transform (DFT)**. This is used in order to convert the signal to the frequency spectrum.

4. **Compute Mel-spectrum**. This is computed by passing the signal through the Mel-filter bank (F-bank), which is a set of triangular band-pass filters over the frequency spectrum. Notably, the Mel-scale is constructed to represent the way human ears would interpret a sound signal, meaning the filters at higher frequencies have a larger bandwidth. This is because differences in frequency at higher frequencies are less noticeable by humans.

5. **Discrete cosine transform (DCT)**. This step creates the final uncorrelated coefficients called Mel frequency cepstral coefficients (MFCCs).

After step four, so called F-bank features have been created, however they can suffer from a high degree of correlation, meaning some networks can't handle them very well. The solution to this is to apply a DCT (i.e. the fifth step) to the F-bank features, creating uncorrelated coefficients (MFCCs).

Historically, the conversion of F-bank features into MFCCs has in many instances been necessary to remove correlations. However, the recent trend is to use the F-bank features directly and let the network deal with the correlation issues.

1.1.3 Speaker identification and verification

At the time of publishing, Desplanques et al. (2020) set the state of the art in speaker verification with their new network (ECAPA-TDNN), an update on the previously successful x-vector network. This is still the most successful network that uses some form of feature engineering, in this case 80-dimensional F-bank features.

Most of the cutting-edge research within creating embeddings from voice focuses on creating generally applicable embeddings. In October 2021, Yang et al. (2021) introduced the SUPERB-benchmark, which compares the performance of different voice embedder networks 13 different tasks. The speaker-related tasks of this benchmark are speaker identification (classifying the specific speaker, using VoxCeleb), speaker verification (determining if two voices are from the same speaker) and speaker diarization (separating multiple speakers from the same sound). These are of course not exactly gender and age classification, however, it is expected that the embedders that provide the most useful embeddings for the tasks described above are also the best candidates for this thesis.

At the time of writing, WavLM-large is the network that tops the SUPERB leaderboard on all speaker-related tasks. This is a network created by Microsoft's UniSpeech project. The same project hosts the UniSpeechSAT-large model, which boasts a very similar performance within speaker-related tasks, however is not listed on the SUPERB leaderboard for unknown reasons. Chen et al. (2021b) describe how SAT stands for "Speaker Aware pre-Training" which alludes to the fact that speaker information is supposed to be saved through the unsupervised pre-training, not surprisingly making this model good at speaker-related tasks.

In Microsoft's corresponding speaker verification paper, Chen et al. (2021c) describe that ECAPA-TDNN (using F-bank features) is the previous state of the art in speaker verification, which they improve upon by funneling the embeddings of UniSpeechSAT-large into a small ECAPA-TDNN, see Figure 1.1. Specifically, a trainable weighted average is taken of all the hidden layers in the UniSpeech embedder. This ensures information from all layers can be used in the ECAPA-TDNN model. The results they achieve are presumably better than if using a simple MLP on top of the UniSpeechSAT embeddings.

The authors noted that earlier layers in the self-supervised pre-trained model are more associated with speaker attributes, whereas the later layers are more associated with the linguistic content. Hence, it is not surprising that a weighted average outperforms simply taking the final layer for speaker-related tasks.

1.2 Background

1.2.1 ECAPA-TDNN

Desplanques et al. (2020) proposed the ECAPA-TDNN architecture. It is an improved version of the previously existing x-vector architecture proposed by Snyder et al. (2018).

As an input, the network takes 80-dimensional F-bank features of the recording, i.e. a two-dimensional input of 80 channels with a certain number of temporal frames, depending on the length of the recording. The network itself is in the basic sense a stack of 1-dimensional convolutions, ReLU nonlinearities, and batch normalizations, with some residual connections. This is followed by a statistical pooling layer, and a final dense layer. This creates a

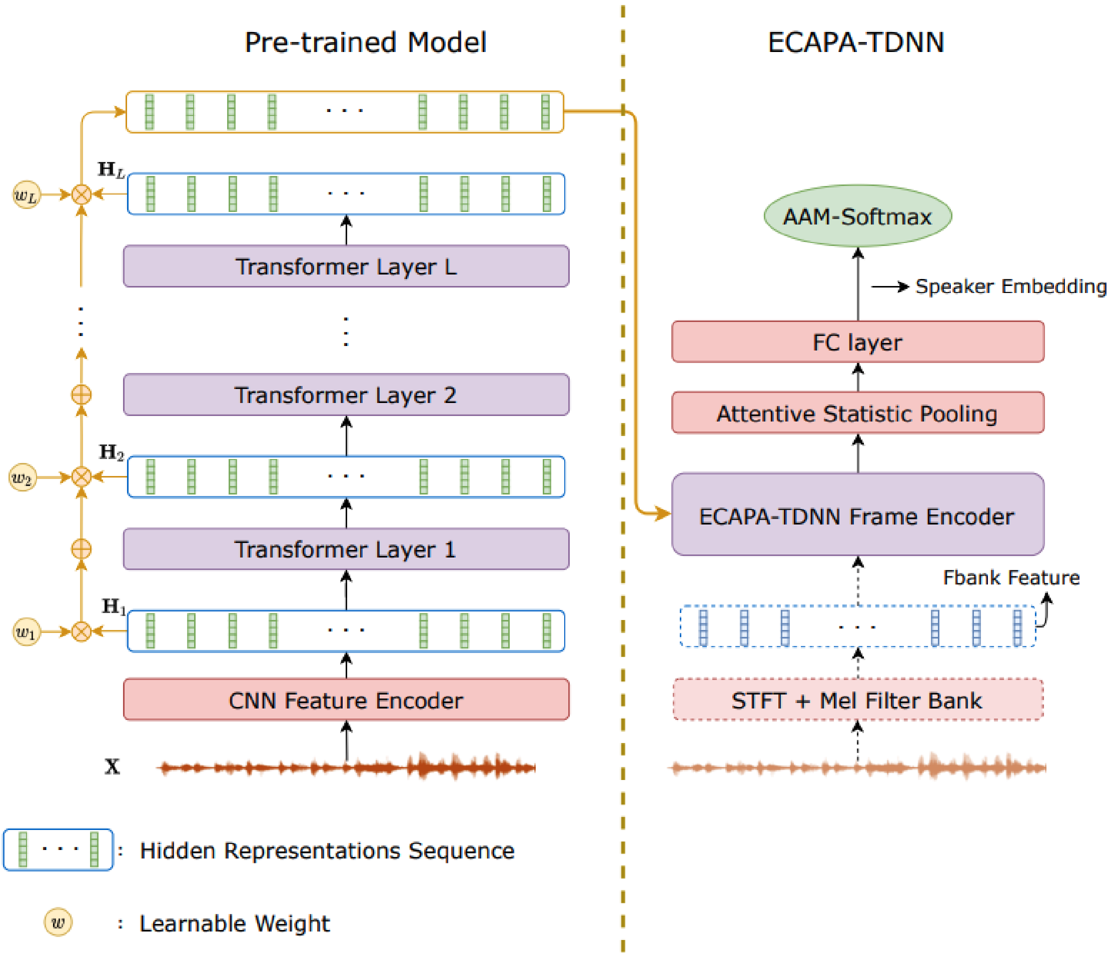


Figure 1.1: The architecture used to get the best performance within speaker identification. Information is passed from all hidden layers of a self-supervised pre-trained model such as UniSpeech-SAT, replacing the F-bank features otherwise used in the ECAPA-TDNN. After (Chen et al., 2021c)

fixed-sized embedding of the variable size (in the temporal dimension) input.

Desplanques et al. (2020) wanted to extend the attention mechanism used in the statistical pooling layer to also include the channel dimension. Previously, soft self-attention had only been used to decide which (time-)frames are the most important. The authors suspected this left some information unnoticed as different speaker characteristics might be more easily detected at different sets of frames. They hence adopted the attentive statistics pooling proposed by Okabe et al. (2018), but modified it to also be channel-dependent, i.e. containing a complete set of temporal attention weights for each channel independently.

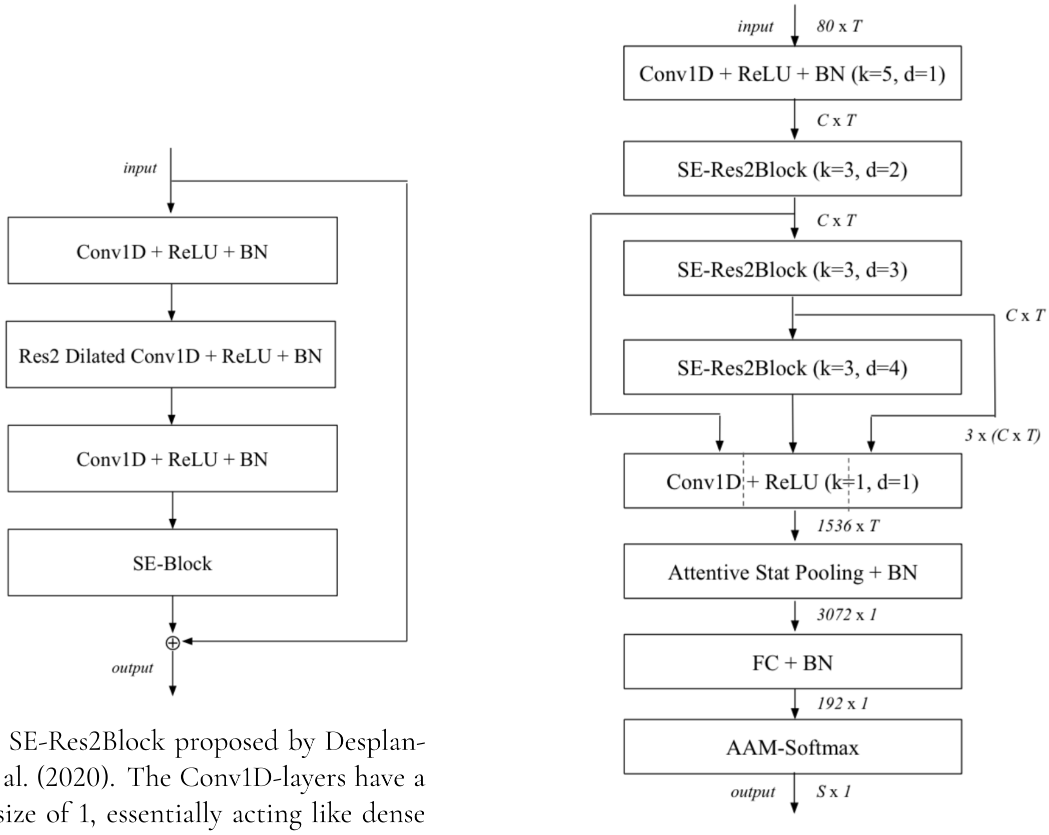
Furthermore, in order to make the attention mechanism more adaptable to global properties of the recording, such as noise and recording conditions, the global mean and standard deviation of each channel are concatenated to the input to the statistical pooling layer described above.

In order to further improve the attention mechanism in the network, Desplanques et al. (2020) also introduce the computer vision inspired 1-dimensional squeeze-excitation blocks

(SE-blocks). The squeeze operation merely calculates the so-called descriptor of each channel by mean-pooling over the temporal dimension. The excitation operation however, uses trainable weights and the channel descriptors to figure out a number between 0 and 1 for each channel. These numbers are subsequently channel-wise multiplied by the input of the block, producing the output, a channel-wise reweighting.

These SE-blocks are then incorporated into the x-vector architecture in a way that is not obstructing the benefits of residual connections and not raising the total parameter count too much. This involves a dense layer to reduce the feature dimension before the dilated convolutional layer, as well as another dense layer to restore the dimensionality afterwards, before applying the SE-Block described above. This results in the SE-Res2Block shown in Figure 1.2a.

Finally, the authors use what they call Multi-layer Feature Aggregation (MFA) in order to utilize outputs from all the SE-Res2Blocks in the networks, arguing that some important information about the speaker might be hidden in the earlier layers. These outputs are merely concatenated and sent to a dense layer that generates the features to feed into the statistical pooling layer. This is represented by the skip connections shown in Figure 1.2b.



(a) The SE-Res2Block proposed by Desplanques et al. (2020). The Conv1D-layers have a kernel size of 1, essentially acting like dense layers, whereas the dilated convolution has a kernel size of 3 and a dilation factor of 2, 3 or 4.

(b) The full architecture of the ECAPA-TDNN network proposed by Desplanques et al. (2020).

1.2.2 HuBERT and HuBERT-based models

Hsu et al. (2021) proposed a model called HuBERT, or Hidden-Unit BERT, to combat the biggest challenges in self-supervised speech representation learning. The backbone of HuBERT is a transformer encoder model, where the central idea is to learn speech representations by an iterative re-clustering and re-training process. The authors state that after only two iterations of clustering, HuBERT improves upon, or at least matches, the previous state of the art.

Every iteration starts with an offline clustering step of each indexed time frame $t : 1 \rightarrow T$ in the input signal. To initialize the clustering, the first iteration clusters the MFCCs of the input utterance. The cluster center of each frame is indexed with a pseudo-label z_t, which we denote as $Z = \{ z_t \}_{t=1}^{T}$.

Next, we extract a feature sequence $X = \{ x_t \}_{t=1}^{T}$ from the utterance with the CNN encoder and mask a subset of the features. We denote $M \subset \{ t \}_{t=1}^{T}$ as the randomly selected subset of indices to mask the corresponding features in X. We can write $\tilde{X} = r(X, T_m)$ as the corrupted feature sequence, where each $\{ x_t \mid x_t \in X, t \in M\}$ is replaced with a random-initialized mask embedding.

The transformer model $f_t(\cdot)$ is then trained to predict the correct labels $\{ z_t \mid z_t \in Z, t \in M\}$ of the masked features, based solely on the context of the unmasked features in the corrupted feature sequence $\tilde{X}$. The model is trained with a mask prediction loss, i.e. a training criterion based on the masked indices $t \in M$, specifically the cross-entropy loss $\mathcal{L} = \sum_{t \in M} \log f(z_t \mid \tilde{X}, t)$.

For the following iterations, the model will cluster the representations generated by the previous iteration of the model, not the MFCCs. Thus, throughout the iterations HuBERT will improve on extracting acoustic features from the continuous inputs, while also refining a language model to predict information based only on context.

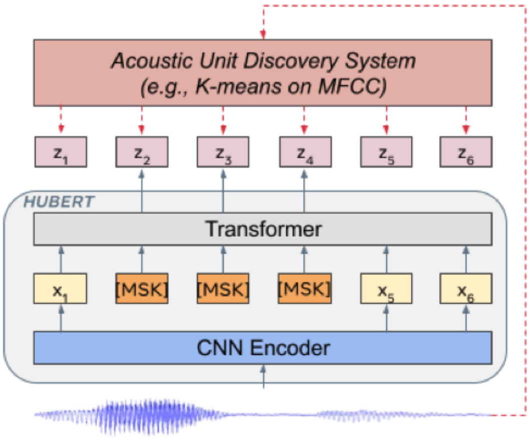


Figure 1.3: HuBERT model architecture, showcasing the transformer and encoder backbone with masked prediction of the clustered pseudo-labels.

As previously mentioned, Yang et al. (2021) created the speech processing universal performance benchmark (SUPERB) to evaluate different model's performances on various speech tasks. HuBERT is primarily trained for automatic speech recognition (ASR) but as the SUPERB leaderboard shows, HuBERT achieves a very respectable performance in multiple speech processing tasks. Because of this, other models have recently used HuBERT as a foundation to build upon, where learning more versatile representations is the main focus.

That is, extending HuBERT and incorporating different learning strategies to carry out tasks beyond ASR and thus produce better speech representations across all tasks. We use two of these models, namely UniSpeechSAT (Chen et al., 2021b) and WavLM (Chen et al., 2021a).

UniSpeech-SAT

UniSpeech-SAT, or **Uni**versal **Speech** representation learning with **S**peaker **A**ware pre-**T**raining, is a model proposed by Chen et al. (2021b), which applies *utterance-wise contrastive learning* and *utterance mixing augmentation* on top of HuBERT. Both techniques improve the representations differently, where the former enhances single speaker information and the latter multi-speaker. Thus, we get a more well-rounded model that can handle different speech processing tasks. Figure 1.4 shows the model architecture together with a schematic view of the two learning techniques.

The main idea of utterance-wise contrastive learning is to teach the model how to produce latent representations that have a low cosine similarity between different speakers. Suppose that we have a latent representation $L^b = \{ l_t^b \}_{t=1}^T$ of a feature sequence for the b-th utterance in a batch. The latent representation L^b is then passed to a quantization module and discretized to a finite set of speech representations $Q^b = \{ q_t^b \}_{t=1}^T$. Now, for each masked latent representation $l_t^b, t \in T_m$ in the b-th utterance, we group all masked quantized speech representations $\tilde{Q}^b = \{ q_t \mid q_t \in Q^b, t \in T_m\}$ from the same utterance. We also include candidate masked speech representations uniformly sampled from other utterances in the same batch $\tilde{Q} = \cup_{b=1}^B \tilde{Q}^b$. The model is then trained to identify the real quantized representations from the candidate ones.

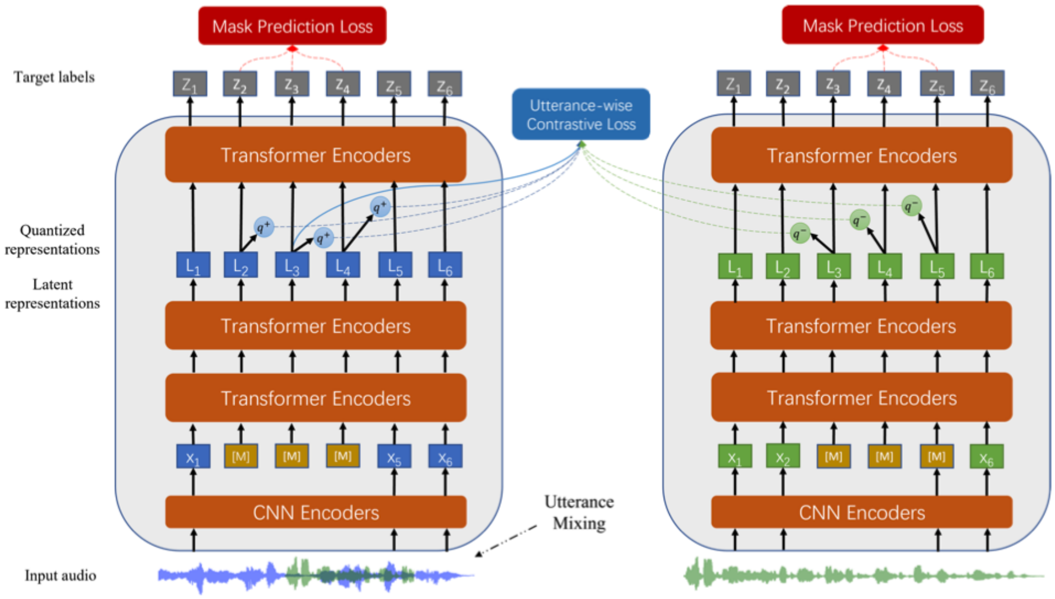


Figure 1.4: Schematic of the model architecture and learning techniques of the UniSpeech-SAT model by Chen et al. (2021b)

In tasks such as speaker diarization, where the goal is to distinguish between different speakers, problems arise when only single-speaker data is available. By using an utterance mixing strategy, a batch of single-speaker utterances can be augmented to simulate multi-speaker data. Given a set of N utterances $U = \{ u_i \}_{i=1}^{N}$, S utterances $\{ \tilde{u}_j \}_{j=1}^{S} \subset U$ are randomly selected for augmentation. For each of these utterances $\tilde{u}$, a random utterance $u \in U$ is selected. A random part of u is then mixed into a random region of $\tilde{u}$. From the mixed audio, the model is trained to extract representations and predict the content information of only the main speaker, not both. Therefore, the model will be forced to learn how to distinguish the main speaker from the mixed speaker in the convoluted audio and thus also improve its performance on multi-speaker tasks.

WavLM

Chen et al. (2021a) introduced the WavLM model in order to explore self-supervised learning (SSL) for speech processing tasks other than automatic speech recognition (ASR). WavLM achieves state-of-the-art performance on the SUPERB benchmark by utilizing denoising in addition to the HuBERT's masked speech prediction during pre-training. The denoising creates robustness to different environments and improves on the extraction of speaker information for non-ASR tasks.

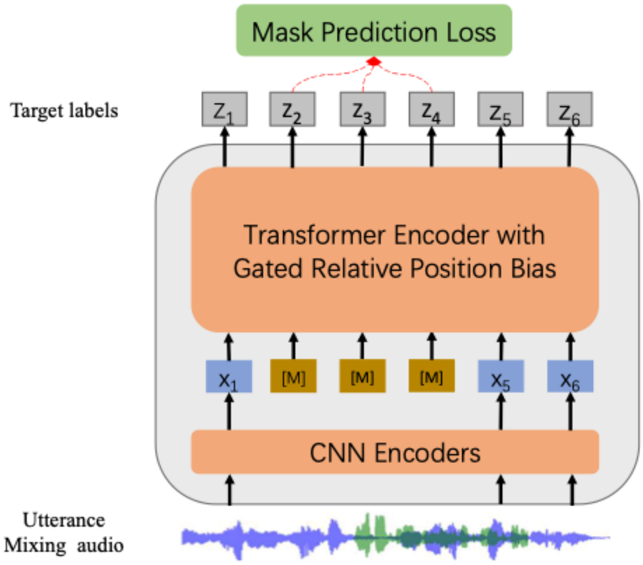


Figure 1.5: The WavLM model architecture with a schematic view of utterance mixing and masked prediction loss.

Similarly to UniSpeech-SAT, an utterance mixing of the inputs is done as a pre-processing step. In each batch, a random selection of utterances is mixed with either randomly selected noise or a random region of a secondary utterance in the batch. This creates both multi-speaker data and noisy speech in the self-supervised pre-training step.

WavLM uses HuBERT's masked prediction loss which implicitly trains the model to denoise and identify the main speaker in the noisy or overlapped input utterances. The masked prediction also trains the model to predict content information from the main speaker's utterance based on context. Therefore, the model learns to first denoise the speech in order to make a better prediction. We can see how WavLM learns to extract speech representations relevant for non-ASR tasks as it implicitly learns to denoise mixed audio and find the most relevant speaker information.

1.2.3 Wav2Vec 2.0 and XLS-R

Babu et al. (2021) presented a large-scale model for cross-lingual speech representations, called XLS-R. The model is a wav2vec 2.0 (Baevski et al., 2020) based model, using self-supervised pre-trained with 0.3B, 1B and 2B parameters on public speech data in 128 languages. XLS-R improves on many benchmarks as it is fine-tuned for tasks regarding language, such as speech recognition, language identification, and speech translation. Additionally, it performs very well on speaker identification according to the SUPERB benchmark (Yang et al., 2021).

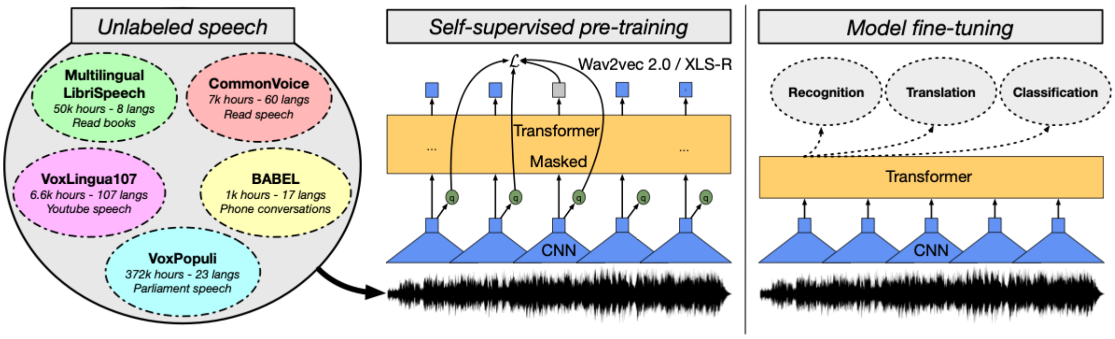


Figure 1.6: Schematic view of the XLS-R architecture with it's self-supervised pre-training on unlabeled multilingual speech as well as the fine-tuning for specific speech processing tasks.

The wav2vec 2.0 model consists of a CNN encoder $f : X \to Z$ mapping raw audio X to a latent representations $Z = \{ z_t \}_{t=1}^{T}$. Every z_t represents 25ms of audio, while the encoder strides the audio with 20ms, creating a slight overlap, which can be seen in Figure 1.6. A transformer $g : Z \to C$ then creates context representations $C = \{ c_t \}_{t=1}^{T}$ from the latent representations. The latent representations from the encoder are also discretized into a set of speech representations $q = \{ q_t \}_{t=1}^{T}$ using a quantization module $h : Z \to Q$.

The wav2vec 2.0 model adopts a learning objective similar to HuBERT (Hsu et al., 2021), where a certain proportion of latent representations are masked before being fed into the transformer. We denote this as $\{ \tilde{z}_t \mid \tilde{z}_t \in Z, t \in M\}$, where M is the set of masked indices.

The model is then tasked to identify the corresponding quantized speech representations $\{ \tilde{q}_t \mid \tilde{q}_t \in Q, t \in M\}$ of the masked time steps based on the context representations $\{ \tilde{c}_t \mid \tilde{c}_t \in C, t \notin M\}$. It does this with a selection out of a set combined with distractors, i.e. in a set with quantized candidate representations $\{ q_t \mid q_t \in Q, t \notin M\}$. This can be seen in the middle of Figure 1.6, where the contrastive loss used to train the model is based on multiple quantized representations as well as a masked time step. Thus, the model learns to extract features for intent classification as the task is context driven and only learns features related to speaker information implicitly.

1.2.4 Ordinal classification

In machine learning, we often make the distinction between classification and regression problems. Regression produces a continuous numerical output, while classification is discrete. Additionally, in regression problems, the targets are ordered and a prediction will be a certain distance from its corresponding target, compared to classification where a prediction is either correct or incorrect.

For age estimation there is a clear ordering in the data. A 60 year old person likely sounds more similar to a 70 year old than a 20 year old. Thus, regression seems appropriate. However, if we have age groups rather than exact ages, we still have an ordering, but our problem has transformed into a classification problem. To utilize the ordering of our data in a classification model, we can apply something called ordinal classification.

Ordinal classification takes the classes' order into account by modifying the network and loss function. The output layer uses the same number of nodes as classes, but also sigmoid activation, rather than the softmax activation commonly used in regular multi-class classification. The prediction is then obtained by looking at how many of the first/lowest output nodes are above 0.5. For example, given five classes, the output activations for a specific input could be: [0.9, 0.8, 0.6, 0.2, 0.1]. This would amount to a prediction of the third class, as the first three output nodes are above 0.5.

The loss function can be chosen in a couple of different ways. In this thesis, we chose to utilize a mean-squared-error (MSE) loss for each individual output node and simply sum these losses to get the total loss of a given voice clip.

We utilized the *mean absolute class error* (MACE) metric in order to measure how good an ordinal classification is. The absolute class error is simply the number of classes separating the prediction from the truth.

In order to put equal importance on all ages although the datasets are skewed, we calculated the MACE for each class respectively and we used the macro average over all the classes as a final metric.

1.3 Problem statement

This thesis aims to investigate the applicability of machine learning models to predict the gender and the age of a person, solely given a recording of the person's voice, in the setting of conversational applications. In order to do this, we will try to answer three main questions:

1. Can machine learning models classify the gender and estimate the age of a person within a useful error margin, using only a clip of that person's voice?

2. Could such models be used in multilingual settings and perform well on unseen languages, i.e. be language agnostic?

3. Could such models be used in a real-time manner, and thus be applicable for real-time conversational applications?

Chapter 2

Dataset

Speech-enabled applications using machine learning require large amounts of data and normally this data is owned and kept private by companies. However, in this thesis we use Mozilla Common Voice, one of the few large public datasets, as well as TIMIT, a much smaller yet commonly used English dataset. We explain how we created datasets suited for our tasks by selecting portions of the Common Voice corpus, and we outline their properties in an exploratory data analysis. A very minimal exploration of TIMIT is also provided.

2.1 Mozilla Common Voice

The Mozilla Common Voice project by Ardila et al. (2020) is a initiative to help offset the lack of public data. By introducing a large, public, open-source voice database that's easily accessible, advancements in machine-learning based speech technology are encouraged.

The Common Voice database is built through crowd-sourcing. Each voice clip in the database is recorded and donated by a voluntary contributor who reads from a bank of sentences in a language of their choice. The recorded voice clip is then queued for validation, where other users will listen to the recording and verify that the sentence is read correctly. If a clip gets two or more confirming votes, it is accepted into the final corpus. Conversely, two or more disapproving votes lands the clip in the so-called clip graveyard.

The database entries include an MP3-file with the recorded voice as well as a text file with the corresponding sentence. Many entries also have demographic metadata including age, sex, and accent. The age is not exact, but rather labeled in 10-year intervals ranging from teens to nineties. At the time of writing, the latest version of the dataset is *Common Voice Corpus 7.0* which consists of 11,192 validated hours in 76 different languages.

2.2 The Common Voice XL Corpus

We used Mozilla Common Voice Corpus 7.0 by Ardila et al. (2020) as a basis for building our different datasets. The corpus is divided into 76 different languages and for each language there exists a validated dataset which contain clean data, approved by the Common Voice community. We concatenated all these validated datasets across the whole corpus to build one unified pool of voice clips.

Due to the lack of demographic data in the corpus, we could not use a large portion of it. Instead, we selected a subset of the corpus, where both the age and the gender of the speaker was documented. Subsequently, this means that we discarded data with only one of those labels, even though it would be sufficient for the respective age or gender classification tasks. However, the simplicity of having the same collection of voice clips for all datasets across both tasks was prioritized.

The speakers of the corpus also have a varying amount of entries due to the open nature of the project. To remove any bias towards more prevalent speakers, this was regularized, and a maximum of five clips per unique speaker was set for each language. We selected clips from each speaker based on their occurring order in the corpus and thus the clips following the first five examples were disregarded. The clip length was also limited to 15 seconds and instead of discarding clips exceeding this limit, we cut the clips and only used the first 15 seconds.

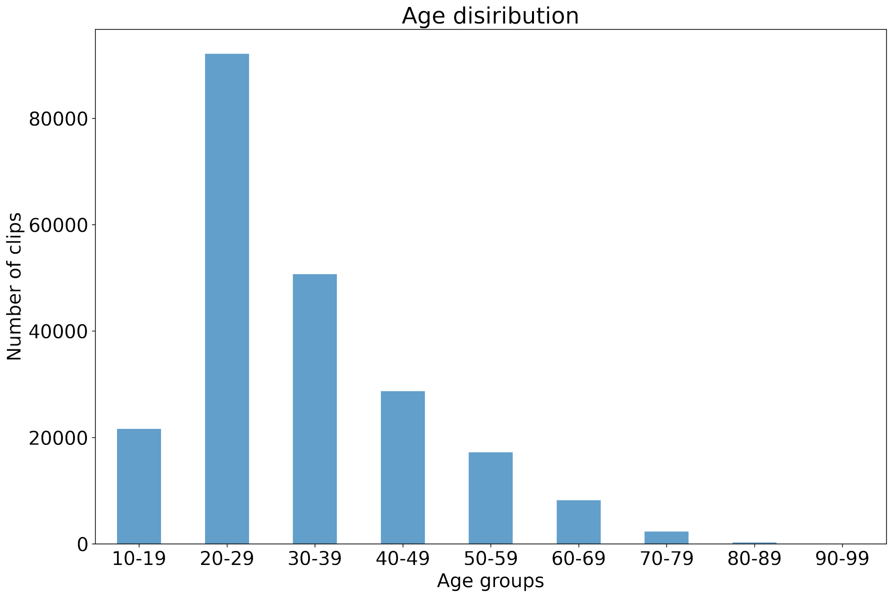


Figure 2.1: The age distribution in the Common Voice XL corpus showing the number of clips per age group. The mean age is 33.3 years with a standard deviation of 13.3 years.

Due to the large original size of the corpus, the reduced version still contains a lot of data. We call the reduced version of the Common Voice Corpus 7.0 with full demographic metadata *Common Voice XL* or *CVXL*. It contains 74 different languages, 43,255 unique speakers, 318 hours and 221,211 clips of recorded voice data and require over 70 GB of space if stored at its full quality of 16 kHz sample rate.

2.2.1 Exploratory data analysis

We're now going to look deeper into CVXL and do an exploratory data analysis to get a better understanding of the corpus. The first aspect we'll consider is the gender distribution. The corpus consists of 168,481 (76.2%) male voice clips and consequently 52,730 (23.8%) female voice clips. With regard to speakers, the voice clips relate to 33,136 male speakers and 10,119 female speakers. It's clear that we have strong male bias in our corpus.

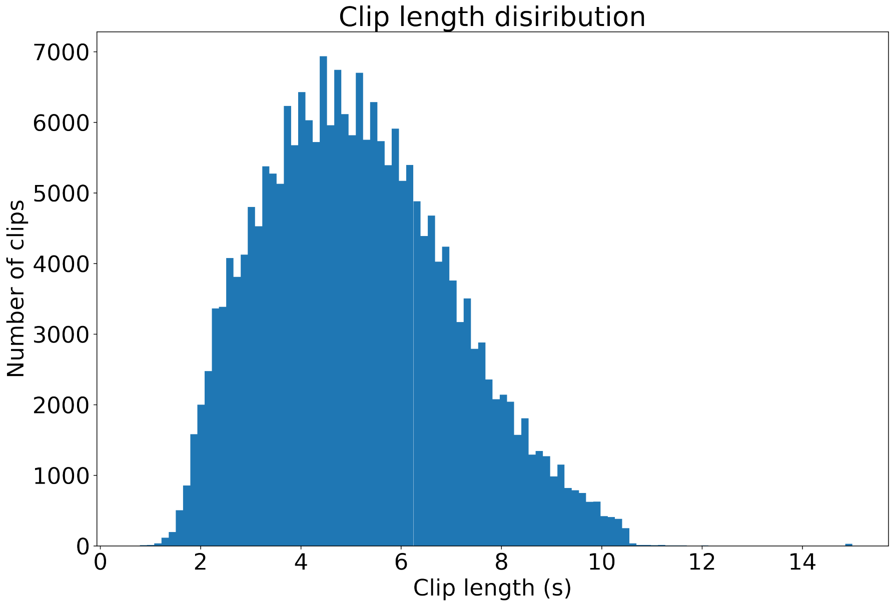


Figure 2.2: A histogram of the clip length distribution in the Common Voice XL corpus with a 100 bins. The mean clip length is 5.2 seconds with a standard deviation of 1.9 seconds.

Figure 2.1 shows the distribution of voice clips with respect to the age groups. By assuming each speaker's age is in the middle of their respective age group, i.e. regarding a speaker in their forties as 45 years old, we can calculate a mean. The mean age in the corpus is 33.3 years with a standard deviation of 13.3 years, which is significantly lower than the median age of 50. It's evident that we have an imbalance and there is greater support for speakers in their twenties, which represent 40.6% of the data, and progressively less as the age increases. Teens also have low support and stand for only 9% of the voice clips, matching the support for speakers in their fifties with 8.4% of the voice clips. The older age groups represent a substantially small amount of data with respect to the whole corpus.

Speakers in their seventies, eighties, and nineties together represent a mere 1.2% of the recorded voice clips and it's therefore probable that a model would have difficulties in distinguishing between these age groups. Even if it was possible, the data lacks statistical significance and do not portray the real world age distribution sufficiently well as there are too few examples across these older age groups, especially for the 80-100 years range. From a practical standpoint, there is also less applicable value to being able to distinguish speakers between these higher age groups even if the data could support such a distinction. There is thus a basis for unifying the subset of clips from speakers in their seventies, eighties and nineties as one 70+ age group and regarding it as one class.

Another aspect of the CVXL corpus is the clip length, which as already mentioned, is limited to 15 seconds. Figure 2.2 shows the distribution of clip lengths across the corpus which has a mean of 5.2 seconds and a standard deviation of 1.9 seconds. It's evident that almost all of the voice clips are less than 11 seconds long, in fact 99.95%, but we can see a small uptick at 15 seconds, where all clips exceeding the 15 seconds limit have grouped. The outlying voice clips do not feature speakers reading unusually long sentences but rather speakers who fail to stop recording and subsequently pollute the clip with pure background noise for up to 10 seconds.

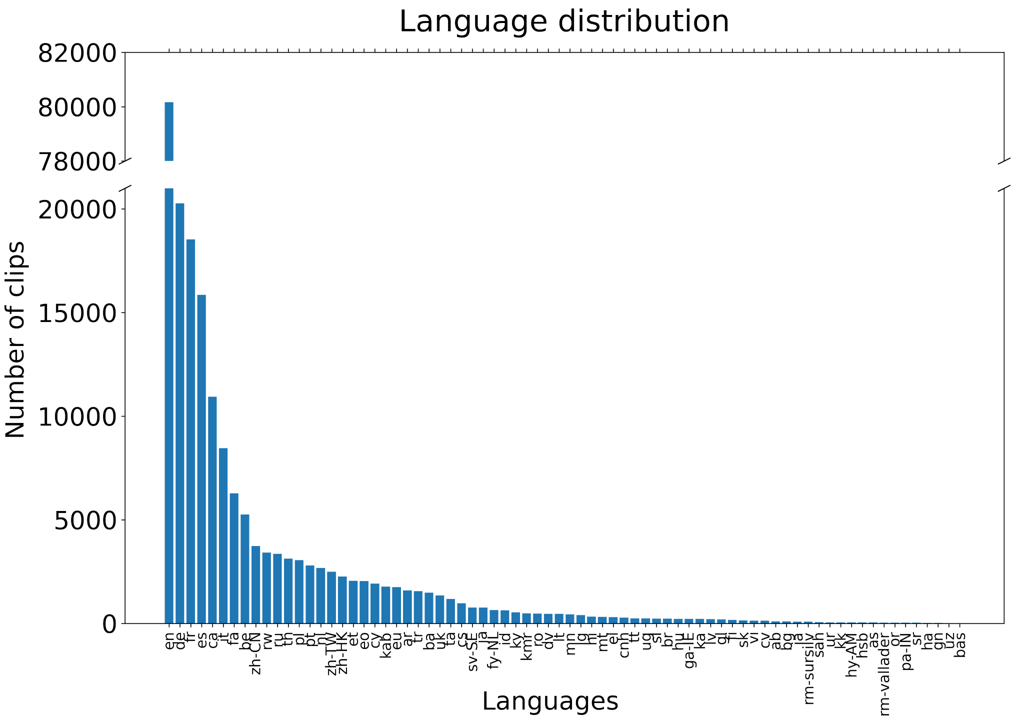


Figure 2.3: The language distribution in the Common Voice XL corpus with regard to the number of clips. The languages are sorted in descending order and displayed in their ISO format.

The final attribute of the corpus that needs to be explored is language. Here, we also see a big difference between languages ranging from 0.3 minutes of recorded speech in Basaa (bas), a Bantu language, to 6,882 minutes, or 115 hours, of English (en) speech. Figure 2.3 shows the distribution across all 74 languages and Table A.3 shows the details of it. Due to this large difference, merely splitting the data into train, validation and test sets without thought would result in a heavy language bias in both training and evaluation. This brings us to the next part of CVXL, namely the four datasets we created.

2.2.2 The four datasets

With the CVXL corpus built and explored, the next step is to split the data into training, validation, and test sets. Since we are to produce language-agnostic models, we need to prove that our models can perform sufficiently well regardless of the language spoken. We will therefore not use all languages in training, but rather a subset of them, and use the rest for validating and testing performance on unseen languages.

From the CVXL corpus, we constructed four different datasets: *Common Voice XL English (CVXL Eng)*, *Common Voice XL 1 Biggest (CVXL 1b)*, *Common Voice XL 5 Biggest (CVXL 5b)*, *Common Voice XL 20 Biggest (CVXL 20b)*, where the number corresponds to the number of training languages. Each dataset uses the specified number biggest languages for training and the rest for evaluation, except for the English dataset, which uses English for training and evaluation. The training and test languages used in each dataset can be seen in Table 2.1, together with their respective size, amount of speakers and the training set split of the whole CVXL corpus.

Table 2.1: The four datasets and their respective training and testing languages written in ISO format. Due to the large number of testing languages, they were denoted as the complement to the training languages. The total size in hours, number of speakers and the training set split of the whole CVXL corpus is also shown.

Dataset	Train languages (T_L)	Test languages*	Size (h)	Speakers	Split (%)
English	en	en	115	17,687	34.8
1 Biggest	en	$T^c_{L\_1b}$	120	19,737	36.2
5 Biggest	ca, de, en, es, fr	$T^c_{L\_5b}$	214	31,583	65.9
20 Biggest	be, ca, cy, de, en, eo, es, et, fa, fr, it, nl, pl, pt, ru, rw, th, zh-CN, zh-HK, zh-TW	$T^c_{L\_20b}$	292	40,730	89.8

* $T^c_{L\_x}$ denotes the complement to the training languages for dataset x, i.e. all languages except for the ones used in training. This is due to the large number of testing languages.

When fewer languages are used in the training set, it also represents a smaller portion of the corpus, making the validation and test sets unnecessarily large. In efforts to speed up the training process, as well as improving language balance in the evaluation sets, we did not use all available data in the validation and test sets. The evaluation data was incrementally reduced to a desired size by capping the number of clips on each age group to 500 and maximizing language diversity for each age group individually. When the number of clips to extract for each language-age combination was decided, maximal speaker diversity was also applied when selecting the speakers to put in the evaluation set. Finally, the evaluation set was split into a validation and a test set by applying a random 50/50 split of the speakers within each age group.

In both CVXL 5b and CVXL 20b we have a diversity of languages. The CVXL 5b training set only contains Indo-European languages, i.e. English, German, Spanish, Catalan, and French. This is an interesting division to see how well western training languages can be used to learn globally applicable speech representations.

In CVXL 20b we include additional Indo-European languages such as Russian, Belarussian, Polish, Dutch, Italian, Welsh and Portuguese, but we also add languages from other families. We have Kinyarwanda from the Atlantic-Congo family in Africa, Estonian from Uralic family in northern Eurasia, Thai from the Kra-Dai family in Asia, Esperanto which is a constructed language and Chinese from the PRC, Hong Kong and Taiwan which belongs to the Sino-Tibetan family. Thus, in CVXL 20b we combined speech from languages covering many central parts of the world, but with a majority of Indo-European data.

We also created an English dataset as a baseline to see how well a model could perform in a monolingual setting. Here, we wanted 500 clips from each age group for the evaluation sets, so the most closely corresponding number of speakers was extracted randomly from each age group, although max 20% in order to retain most of the data for training. Again, the validation and test sets were created by a random 50/50 split of the unique speakers for each age group in the evaluation set.

We did not use gender information when creating the datasets, as initial tests deemed the gender recognition task easier than age detection. We therefore prioritized age balance when creating the evaluation sets.

A detailed view of each dataset with the amount of data for each age group, gender and language can be found in Appendix A.

2.3 TIMIT

The TIMIT corpus by Garofolo et al. (1993) contains English spoken by 630 different individuals with 10 voice clips per speaker. The age distribution can be seen in Figure 2.4. Just as for the Common Voice XL dataset, a heavy bias toward younger people is clearly present. TIMIT is also relatively skewed towards men, as these make up 70% of the dataset.

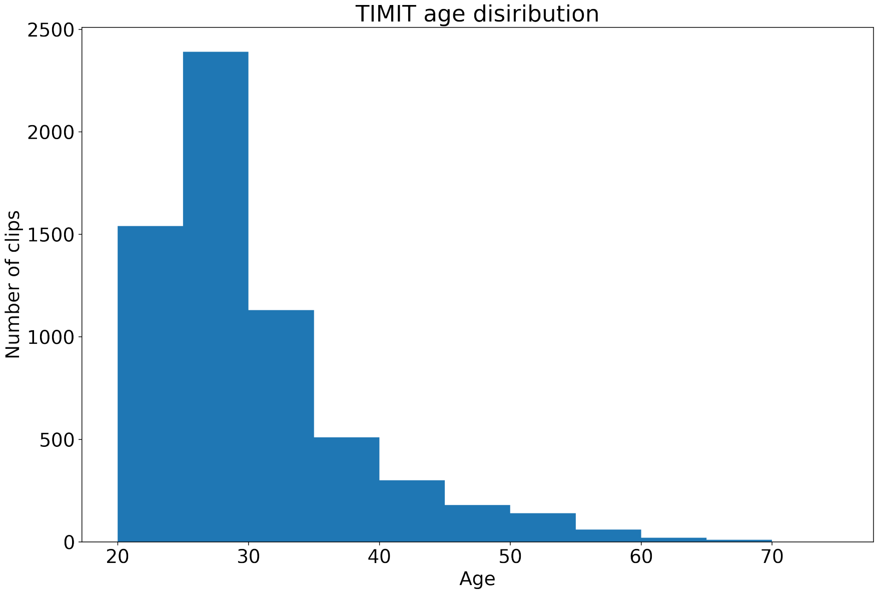


Figure 2.4: The age distribution of the TIMIT dataset. The data is binned into 5-year intervals.

The dataset has a pre-defined train-test split, however we chose to further divide the provided train set into a train and validation set. This was done by randomly taking 20% of the speakers from the provided training set and putting all of their clips into the validation set.

As the TIMIT dataset was created in a laboratory setting, as opposed to the crowdsourced Common Voice, it is significantly cleaner. This will become clear in the results part of this thesis.

Chapter 3
Approach

3.1 Model selection

In Section 1.1.1, we showed that simple acoustic features have potential in gender classification. Therefore, we evaluated MFCC features as a potential lightweight model. The default settings of the MFCC-extractor by SpeechBrain (Ravanelli et al., 2021) was used to extract the features. This meant 20 MFCCs was extracted from 23 F-Bank features. Also derivatives, second derivatives and context parameters were added, resulting in a feature vector of size 660 for each voice clip. Simple MLPs with one hidden layer of size 200 was then used for both tasks. The results from these simple networks will serve as our baseline.

The central notion in this gender and age categorization project was to utilize models pre-trained on related speech processing task and apply transfer learning through a fine-tuning process. Pre-trained models can find more complex relations in data, and consequently solve more complex problems. Thus, pre-trained models should probably outperform any model built from scratch as building a model from scratch requires a huge dataset as well as time and resources that we did not have. This is especially true in age classification as it is a more complex task than gender classification.

We therefore selected a handful of pre-trained models with high performance in single speaker speech processing tasks, such as speaker identification and speaker verification. Both speaker identification and speaker verification require models to capture information about the speaker, which is highly relevant to both age and gender classification. The selection was based on the SUPERB benchmark by Yang et al. (2021), which is a standardized testbed for a comprehensive evaluation of pre-trained models in various speech processing tasks. The models we selected are:

- ECAPA-TDNN
- XLS-R 300M
- UniSpeech-SAT Large

- WavLM Large

ECAPA-TDNN by (Desplanques et al., 2020) was the state of the art in speaker verification at the time of publication, and was only outperformed in this task in the fall of 2021. We believed it still had some potential for gender and age classification, and wanted to include it in this thesis because it is significantly smaller than the other models, and is based on F-bank features as opposed to the audio signal directly. In a way, comparing the results of ECAPA-TDNN to the newer transformer models shows how the transformer technique has changed the quality of audio embeddings.

WavLM (Chen et al., 2021a) and UniSpeech-SAT (Chen et al., 2021b) are similar models, as they both extend the same base model, HuBERT (Hsu et al., 2021). They both top the SUPERB leaderboard with different methods of transforming HuBERT to generate more universally applied speech representations across multiple speech processing tasks. As they achieve state-of-the-art results, they are of interest to apply in gender and age classification as well.

We have also included a Wav2Vec 2.0 (Baevski et al., 2020) based model, trained on multilingual data to produce cross-lingual speech representations, namely XLS-R 300M (Babu et al., 2021). XLS-R 300M shows intriguing qualities in the language-agnostic aspect of our project as it produces even more general speech representations, relevant across different languages. XLS-R is mostly focused on improving on benchmarks for tasks regarding language, such as speech recognition, language identification, and speech translation. However, it still produces quality results on speaker identification with regards to the SUPERB benchmark. XLS-R 300M thus introduces multilingual capabilities while preserving relevant speech representations for our tasks.

All pre-trained models are available through HuggingFace[1] (Wolf et al., 2020) and SpeechBrain[2] (Ravanelli et al., 2021).

3.2 Baseline with pre-trained embeddings

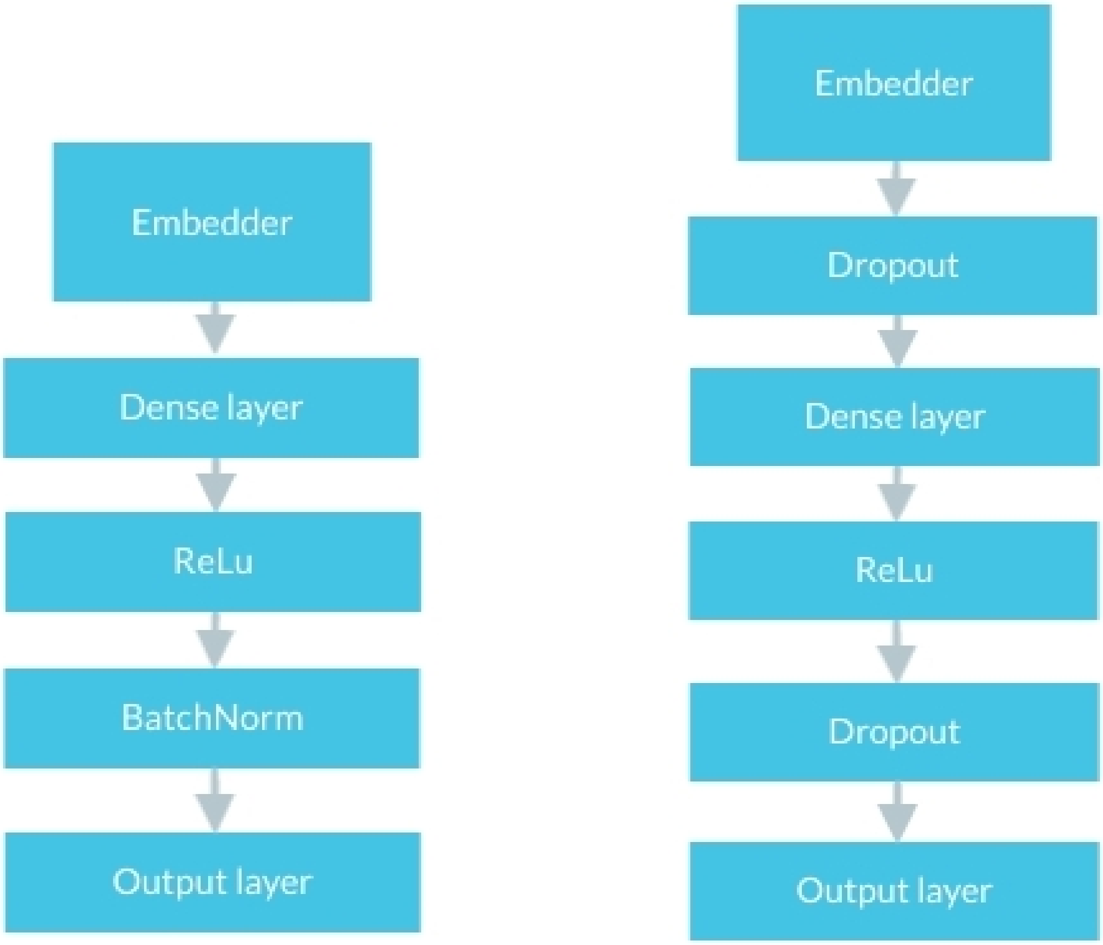
With a variety of models selected, we are to compare and outline their performance on our previously unseen tasks, i.e. gender and age classification, to create a benchmark. Here, the MFCC models will serve as a baseline to compare with the more complex pre-trained models.

We initially utilized the pre-trained models as pure embedders to extract latent representations of the datasets. With the embeddings for each model and dataset, we trained and evaluated different network heads for our two different tasks. The basic schematics of the complete networks can be seen in Figure 3.1.

The classification heads all consisted of neural networks, trained with the Adam optimizer. They all had a single hidden layer of the same size as the embedding dimension, followed by an output layer corresponding to the task and type of data fit. The nonlinearity between the hidden layers and the output layers were all ReLU activations, and for regularization, we either used two dropout layers or a single batch normalization layer. For gender classification, we used a binary classification head with one output node with a sigmoid activation function, trained with binary cross entropy (BCE) loss. In age classification, we tried two different heads:

[1] https://huggingface.co/models

[2] https://speechbrain.github.io



(a) Network architecture with batch norm regularization.

(b) Network architecture with dropout regularization.

Figure 3.1: General network schematics. For the pre-trained results the embedder is frozen during training.

1. A regression head with a single output node with linear activation, trained using absolute error (L1-loss) as a loss function. Here, for the CVXL datasets, the classes assumed an exact age of the average of each age group, e.g. the target of class “fourties” was 45.

2. An ordinal classification head with 7 output nodes corresponding to the 7 classes, all with sigmoid activation. Here, a specific ordinal loss function was used, as described in Section 1.2.4. The loss is the sum of the mean squared errors (MSE) of each output node.

To compare the regression head with the ordinal classification head, the predictions were binned to our 7 age groups. Thus, the regression head serves as a classifier and can be compared with ordinal classification.

The datasets all suffer from heavy imbalance with regards to age and gender, and therefore yield highly biased models during training. We mitigated this by implementing a bin sampling, meaning all classes in a training batch were sampled from a uniform distribution. That is, we’re equally likely to include a highly supported class as a lowly supported class in the training batch. This in return means that not all data is seen equally often by the model during one training epoch. The data from a class with high support is abundant, and thus parts are left unseen, while data from a class with low support is scarce, and therefore recycled.

With our classification heads defined and data imbalance reduced, we obtained a baseline on both tasks for the models across all datasets. Next, we attached the ordinal classification heads to the pre-trained models to begin a fine-tuning process.

3.3 Fine tuning pre-trained models

With a pre-trained baseline across both age and gender classification, we have something to improve upon. As our models are pre-trained for other speech processing tasks, our results are closely related to how relevant the speech representations are to age and gender classification. To improve our results and produce more relevant speech representations, we fine-tuned our models.

The baseline showed us insights in performance across different classifiers and tasks, which consequently led us to the structure of our fine-tuning process. As the baseline for gender classification showed satisfactory result using only the pre-trained embeddings, we proceeded by not fine-tuning any models for gender classification. Instead age estimation was the more challenging task, meaning larger improvements in the embeddings could be achieved using this task for fine-tuning. The results of the age classification networks made it clear that, given the discrete form of our data, ordinal classification was a better fit for the problem than regression. Hence, we proceeded by only using ordinal classification heads for fine-tuning on the age task.

During fine tuning, we adopted the bin sampling method described earlier to combat the age imbalance in the datasets. On top of this, we used data augmentation to create a noisier dataset in the hope of producing more generalized embeddings. The augmentation pipeline was copied from the SpeechBrain-recipe defining the pre-training of the ECAPA-TDNN network. The augmentations consisted of 5 modifications of the clips:

1. Dropping parts of the signal;
2. Modifying the speed +/- 10%;
3. Adding reverb;
4. Adding noise;
5. Adding both reverb and noise.

For the purpose of generalizing the models to recording conditions, all these modified clips were added to the same batch as the original clip. This resulted in an actual batch size 6 times larger than the number of unique clips from the original dataset in each batch.

We fine-tuned the ECAPA-TDNN model following a procedure similar to that of how Ravanelli et al. (2021) had pre-trained the model, only with an age training criteria instead of speaker identification. Similar hyperparameters were adapted as these were assumed to be relevant for this task too. Adam was used for optimization with a weight decay set to

10^{-6}, and the batch size was set to 36. Importantly, due to the low weight decay, two dropout layers of 25% were used in the ordinal classification head, rather than batch normalization. A OneCycle learning rate schedule by Smith and Topin (2019) was adopted, using a peak learning rate of 10^{-4} with linear decay.

We fine-tuned UniSpeech-SAT, WavLM and Wav2Vec2 XLS-R using the HuggingFace's Transformers library. AdamW was the selected optimizer and the batch size was set to 8. We used a linear learning rate schedule starting at 10^{-4} together with a linear warmup of 100 steps. A single batch normalization layer was used in the ordinal classification head as a regularizer.

After the models had been fine tuned on the age task, some improvements were achieved by retraining the classification head, freezing the embedding part of the networks. Additionally, we used the fine-tuned embeddings to retrain classification heads for gender classification, seeing how the fine-tuning on age classification affects the performance. This was done to check the potential of connecting both heads to one embedder, creating one unified model for both age and gender classification.

3.4 TIMIT

The TIMIT dataset was mostly used in order to compare the resulting models to the previously achieved results we found in the literature. For gender classification we maximized the macro average F1-score, however reported the overall accuracy as this is commonly reported. For age detection, we only performed regression, given that the ages are exactly specified and previous results are mainly regression results. In both cases the embedders fine-tuned on the English dataset was used to extract embeddings, and the heads were trained on top of those embeddings.

Chapter 4

Results

4.1 Pre-trained embeddings

We've selected various cutting-edge transformer models in multiple speech processing tasks, as well as a previous state-of-the-art model within speaker verification (ECAPA-TDNN). These models will be used as embedders for an initial gender and age classification benchmark. Additionally, we created a baseline from a simple acoustic feature set of MFCCs.

The MFCCs and the pre-trained embeddings were extracted from our four CVXL datasets, CVXL English, CVXL 1 Biggest, CVXL 5 Biggest and CVXL 20 Biggest in order to gauge the language generalization performance of the models.

4.1.1 Gender

Our gender classification results were achieved by training a simple binary classification head on top of the pre-trained embeddings. Table 4.1 shows the macro average F1-scores for each model across all datasets.

Table 4.1: Macro average F1-score with a binary classification head on the test sets of the different datasets with pre-trained embeddings.

Model	English	1b	5b	20b
MFCC	0.845	0.854	0.870	0.864
ECAPA-TDNN	**0.960**	0.952	**0.954**	0.949
UniSpeech-SAT	0.952	**0.953**	0.948	**0.954**
WavLM	0.957	0.951	0.946	0.947
XLS-R 300M	0.926	0.942	0.942	0.938

All the pre-trained models outperform the simple acoustic features quite clearly. The models also perform very well overall and fairly equal. Only XLS-R 300M performs slightly worse than the other models. This indicates gender classification is an easy enough task for the smaller ECAPA-TDNN to handle just as well as the larger, more complex, transformer models.

When the performances on male and female are compared, it is clear that the imbalance in the datasets, which is roughly 70% male, creates networks better at specifying male voices. In most cases, the male F1-score is 0.03-0.04 higher than the female score. This occurred even though bin-sampling was used to combat the imbalance.

4.1.2 Age

For age classification, we trained both regression and ordinal classification heads on top of the pre-trained embeddings. Table 4.2 shows the regression results for all models across the test datasets, and Table 4.3 shows the ordinal classification results.

In order to evaluate the performance of the regression, both the mean absolute error (MAE) and *(R^2)*-score are shown. By maximizing the *(R^2)*-score, we ensured that our models performed well over all ages, and not just predicting ages close to the mean. This was important given the age imbalance of the datasets, since such a network would still yield a low MAE.

Table 4.2: Mean absolute error (MAE) and coefficient of determination R^2-scores of the age regression on the different datasets with simple acoustic features and pre-trained embeddings.

Model	English		1b		5b		20b	
	(MAE)	*(R^2)*	*(MAE)*	*(R^2)*	*(MAE)*	*(R^2)*	*(MAE)*	*(R^2)*
MFCCs	15.52	0.067	17.48	0.059	16.18	0.094	14.76	0.090
ECAPA-TDNN	9.87	0.579	12.85	0.409	11.43	0.455	10.45	0.435
UniSpeech-SAT	8.84	0.662	12.99	0.39	11.13	**0.496**	**10.42**	**0.468**
WavLM Large	**8.36**	**0.683**	**12.10**	**0.472**	**11.01**	0.48	10.53	0.446
XLS-R 300M	10.72	0.498	12.31	0.469	11,71	0.455	11.05	0.433

The WavLM model performs the best overall, with the UniSpeech-SAT model trailing not far behind. Specifically, the UniSpeech-SAT model seems better when more languages are used for training. We can see that the regression heads perform the best on the English dataset for all models, which indicates a lack of multilingual capabilities in our pre-trained embeddings. However, the regression model improves as it is trained on more multilingual data for all models except the XLS-R 300M. For this model we see less of a difference between the monolingual and multilingual datasets, probably because it was pre-trained for cross-lingual speech representations.

Just like for the regression head, we trained the ordinal classification head for all the models and datasets. In order to compare the two heads, we binned the regression predictions into the same seven age classes used in the ordinal classification. Table 4.3 shows the macro average MACEs across all the datasets and models for the two classification heads.

Table 4.3: Macro average mean absolute class errors with regression and ordinal classification heads, using simple acoustic features and pre-trained embeddings. The regression results were binned into the same age classes used for the classification.

Model	**English**		**1b**		**5b**		**20b**	
	(Ord.)	*(Reg.)*	*(Ord.)*	*(Reg.)*	*(Ord.)*	*(Reg.)*	*(Ord.)*	*(Reg.)*
MFCC	1.608	1.600	1.629	1.642	1.570	1.611	1.606	1.619
ECAPA-TDNN	0.912	0.962	1.186	1.179	1.070	1.080	1.115	1.119
UniSpeech-SAT	0.848	0.836	1.222	1.202	1.046	1.067	**1.079**	1.110
WavLM	0.804	**0.799**	1.125	1.109	**1.002**	1.037	1.081	1.120
XLS-R 300M	1.014	1.027	**1.084**	1.151	1.065	1.099	1.093	1.151

In a monolingual setting on CVXL English dataset, the binned regression head and the ordinal classification head are quite equal. For UniSpeech-SAT and WavLM, the regression head is better with a lower macro average MACE, but for ECAPA-TDNN and XLS-R it's the opposite. Similarly, in CVXL 1b the regression head is only significantly worse for XLS-R. However, for both CVXL 5b and 20b the ordinal classification head outperforms the binned regression head with a lower macro MACE for all the models. Thus, the two heads can be seen as fairly equal in performance, but as the CVXL corpus contains age groups and not exact ages, the ordinal classification head is better suited for our data. We therefore continued by only using the ordinal classification head in the fine tuning stage of this thesis.

4.2 Fine-tuning

With initial results on both age and gender classification using our pre-trained embeddings, the next step is to apply transfer learning. With the pre-trained embeddings, the performance is limited by the applicability of the embeddings to our tasks of age and gender classification. Therefore, we fine-tune the pre-trained models using one of our tasks in order to make the embeddings more tailored to that task. The fine-tuning was only carried out for age classification as our results on gender classification are satisfactory with pre-trained embeddings. Additionally, we only fine-tuned with ordinal classification heads as they proved to be a better fit to the CVXL corpus than regression heads.

Table 4.4: Macro average MACE with an ordinal classification head on the test sets of the different CVXL datasets with fine-tuned embeddings.

Model	**English**	**1b**	**5b**	**20b**
ECAPA-TDNN	0.792	1.040	0.953	1.023
UniSpeech-SAT	0.711	0.983	0.897	0.917
WavLM	**0.682**	**0.910**	**0.857**	0.918
XLS-R 300M	0.784	0.964	0.862	**0.886**

Table 4.4 shows the macro average MACEs for our fine-tuning, where the ordinal clas-

sification heads were re-trained after the fine-tuning process. We can see an improvement in performance across all models and datasets with fine-tuning, when comparing to the results in Table 4.3 with pre-trained embeddings. Evidently, fine-tuning makes the embeddings better suited to predicting the age of the speakers. WavLM stands out as the overall best performing model, particularly when training on fewer languages.

We also investigated if these new, fine-tuned embeddings could be used for improving, or at least match, the results on gender classification. Table 4.5 shows the macro average F1-scores for a binary classification head on the fine-tuned embeddings.

Table 4.5: Macro average F1-score with a binary classification head on the test sets of the different datasets with fine-tuned embeddings.

Model	English	1b	5b	20b
ECAPA-TDNN	0.959	0.955	**0.956**	0.956
UniSpeech-SAT	0.955	0.959	0.954	**0.958**
WavLM	**0.961**	**0.962**	0.951	0.955
XLS-R 300M	0.951	0.959	0.947	0.943

By comparing to the pre-trained results in Table 4.1, we can conclude that even though the fine-tuning was performed on the age task, the resulting embeddings are certainly useful also for the gender classification task. In fact, the F1-scores are slightly better across the board, suggesting that the fine-tuning might have enhanced speaker related attributes in the embeddings.

Again, all models perform adequately, however ECAPA-TDNN stands out as a lightweight alternative to get practically the same results on the gender classification task as the heavier transformer alternatives.

4.2.1 Language agnosticism

Having fine-tuned our models and achieved improved performance with the fine-tuned embeddings, we have provided all necessary information to answer the first question posed in the problem statement. However, in order to answer the second, we need to dive deeper into another important part of our research, the language agnosticism of our models.

Table 4.6: Macro average MACE for the fine-tuned models with an ordinal classification head on the 20b test set. The datasets denote the datasets used for fine-tuning, i.e. CVXL 1b, 5b and 20b, and not testing.

Model	1b	5b	20b
ECAPA-TDNN	1.093	1.004	1.023
UniSpeech-SAT	1.139	0.943	0.917
WavLM	**1.081**	**0.908**	0.918
XLS-R 300M	1.124	0.941	**0.886**

The performance on the CVXL English dataset, i.e. a completely monolingual setting, should be perceived as the optimal performance for these models, at least when trained as we've presented in this thesis.

For the gender task, the performance difference when testing on unseen languages is remarkably small for all models. This suggests that the attributes of a voice pertaining to gender are inherently language agnostic, and thus machine learning models can perform just as well on unseen languages.

However, for the age task, the performance on unseen languages is significantly worse, indicating that the attributes in a voice pertaining to age are in many cases somewhat language dependent.

Comparing the performance on CVXL English and CVXL 1b clearly shows the difference between testing on a seen or unseen language as these two datasets have practically the same training set. For the age task the performance is significantly worse on the 1b dataset, at least indicating that training on a single language is not sufficient for optimal language agnostic performance.

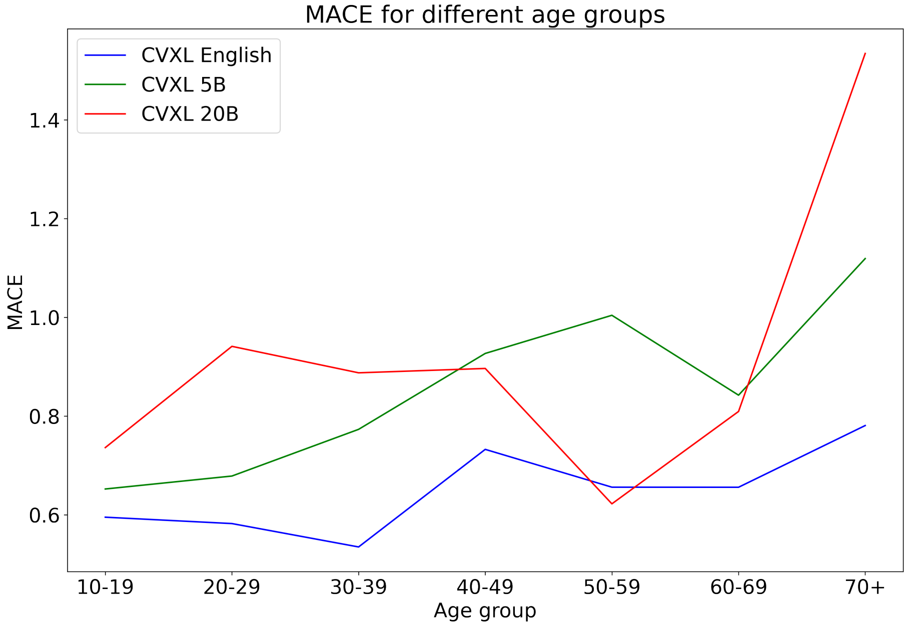


Figure 4.1: The mean absolute class error (MACE) for each age group in both monolingual (CVXL English) and multilingual (CVXL 5b and CVXL 20b) datasets. The MACE-scores are for the WavLM model fine-tuned for each respective dataset.

For the multilingual datasets, we see a natural increase in performance when training on five languages instead of only one, but a loss of performance when increasing that to 20 languages. This can most likely be attributed to that the test set in the 20b dataset is more difficult than for the other datasets, which probably happened purely by chance. In order to shed some more light on this, Table 4.6 contains results on the 20b test set only, using models fine-tuned on other datasets. We do this because the evaluation languages of 20b are unseen by all models. This is a slightly unfair comparison, as the model fine-tuned on the

20b dataset is optimized for performance on the validation set of 20b. However, we think it still clearly shows that training on more languages is always a good thing for performance on unseen languages, although not much is gained beyond 5 training languages.

4.2.2 Age and language imbalance

A natural occurring phenomena when collecting data, is finding imbalances. In Section 2.2.1 we did an extensive data analysis of our CVXL corpus where we found imbalances in all demographic data as neither age, gender nor language follow an uniform distribution.

With bin sampling and augmentation, we've made efforts to mitigate the imbalance, however we're still left with biased models and varying performance over age, gender and language.

Figure 4.1 shows the MACE distribution for an age classification by the WavLM model for different datasets. We can see that the model performs best on younger age groups and the worst on 70+ for all datasets. We also get a more skewed performance in the multilingual datasets than the English monolingual one. The difference between the best and the worst performing class is 0.25 on CVXL English compared to 0.91 for CVXL 20b.

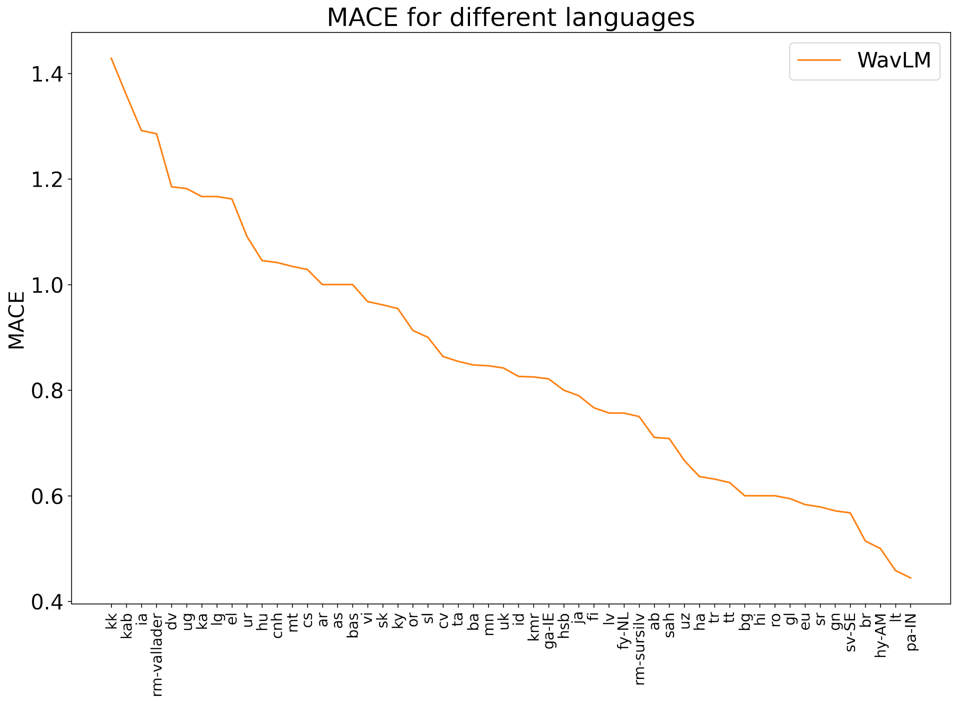


Figure 4.2: The mean absolute class error (MACE) for each test language in CVXL 20b, sorted in descending order. The MACE-scores are for the WavLM model fine-tuned on CVXL 20b.

Bin sampling is supposed to prevent the models from being biased toward the highly supported classes by sampling each class equally often. However, since data from classes of low support is recycled more, our models could overfit to the relatively few examples belonging to those classes. Augmentation is mitigating this effect, however, due to the heavy age imbalance of our Common Voice XL datasets, our models overfit simply due to there being too much data recycling. This effect is especially evident for the 70+ class, where we see a decrease in performance between CVXL 5b and 20b, probably because the training data grows by 36% but with the same support for the 70+ age group.

Similarly, performance across different languages also varies. In Figure 4.2 we can see the MACE for each language in the CVXL 20b dataset. The languages are sorted by their respective MACE score, with languages as Kazakh, Kabyle and Interlingua preforming the worst and Armenian, Lithuanian and Punjabi preforming the best. It is unclear why certain languages yield a better or worse MACE. Neither the amount of examples in the dataset nor the age distribution within the languages seem to correlate in any way with the language specific performances.

4.2.3 TIMIT

To compare our models to previous work and the current state of the art we evaluate our models on the TIMIT dataset. TIMIT is an English dataset containing both the gender and exact age of the its speakers. We will therefore train a binary classification head for gender classification and a regression head for age estimation using embeddings extracted from the TIMIT dataset by models fined-tuned on the CVXL English dataset.

Table 4.7 shows the accuracy of our different models on the TIMIT dataset in gender classification. We can see that both ECAPA-TDNN, WavLM and UniSpeech-SAT beat the previous state of the art model, QuartzNet (Kwasny and Hemmerling, 2021). We achieve the best accuracy of 99.8% with UniSpeech-SAT, improving on the state-of-the-art by 2 ‰.

Table 4.7: Gender classification accuracy on the TIMIT test set using models fine-tuned on the CVXL English dataset.

Model	Accuracy	Macro avg. F1-score
Prior work		
QuartzNet *	0.996	N/A
This work		
ECAPA-TDNN	0.997	**0.997**
XLS-R 300M	0.995	0.994
UniSpeech-SAT	**0.998**	**0.997**
WavLM	0.997	**0.997**

* Model by Kwasny and Hemmerling (2021)

Performance on the TIMIT dataset was significantly better than the one we obtained on our CVXL-English dataset. Gender classification on the TIMIT dataset came in at a macro F1-score of 0.997 for the WavLM, UniSpeech-SAT and ECAPA-TDNN models, compared to our best macro F1-score of 0.961 on CVXL-English with WavLM. This probably mainly has to do with that the TIMIT dataset is cleaner compared to the CVXL corpus. By being built from the open source and crowdsourced CommonVoice project (Ardila et al., 2020), the CVXL corpus is naturally quite noisy.

In Table 4.8 we can see the regression results for the TIMIT dataset. The best result we produced was by the WavLM model and constituted of a MAE of 4.11 for males, 4.44 for females and an R^2-score of 0.54. This can be compared to the previous best result in the literature by Kwasny and Hemmerling (2021), where their QuartzNet model obtained 5.12 MAE for men and 5.29 MAE for females. We beat the previous state of the art with 1.18 years for males and 0.68 years for females with the WavLM model. Thus, we can see the strength of our fine-tuned models compared to previous models in the literature.

Furthermore, we believe optimizing the R2-score, rather than the MAE, creates a better regression in an unbalanced setting, as it is more evenly accurate along all ages. However, Kwasny and Hemmerling (2021) does not provide the R^2-score of their results.

Table 4.8: Age regression on the TIMIT test set using models fine-tuned on the CVXL English dataset.

Model	**MAE** *(Male)*	*(Female)*	*(Total)*	**R2-score**
Prior work				
QuartzNet *	5.12	5.29	5.17	N/A
This work				
ECAPA-TDNN	5.00	5.04	5.01	0.364
XLS-R 300M	4.82	5.16	4.94	0.409
UniSpeech-SAT	4.28	4.54	4.36	**0.541**
WavLM	**4.11**	**4.44**	**4.22**	0.540

* Model by Kwasny and Hemmerling (2021)

4.3 Model applicability

The third question posed in this thesis' problem statement regards the applicability of these models in a real-time setting, such as a voiced conversational application. This section aims to answer that question.

4.3.1 Inference time

Table 4.9: Inference time in ms for 1, 3, 5 and 10 second voice clips. Results averaged over 1000 examples of randomly generated data on a NVIDIA GeForce RTX 3090 GPU

Model	**1s**	**3s**	**5s**	**10s**
ECAPA-TDNN	7.8	7.8	7.5	8.6
UniSpeech-SAT	12.3	12.5	14.0	26.5
WavLM	16.8	18.3	18.8	29.6
XLS-R 300M	12.4	12.7	14.2	26.8

Accuracy is undoubtedly important for any model that is to be used in a real world setting. However, merely high accuracy does not guarantee a models practicality, as it omits an integral part, namely time. Great results hold little practical value if they take too long to produce. Therefore, to attest the practicality of a model, we need to show acceptable accuracy in an acceptable amount of time. That is, the inference time of our model needs to be low enough for the results to be usable in a real-time setting. Table 4.9 and 4.10 show the inference time across our fine-tuned models for voice clips of varying lengths for a GPU and CPU respectively.

Table 4.10: Inference time in ms for 1, 3, 5 and 10 second voice clips. Results averaged over 1000 examples of randomly generated data on a CPU.

Model	**1s**	**3s**	**5s**	**10s**
ECAPA-TDNN	180	404	640	1230
UniSpeech-SAT	315	589	762	1149
WavLM	323	592	774	1206
XLS-R 300M	320	570	752	1248

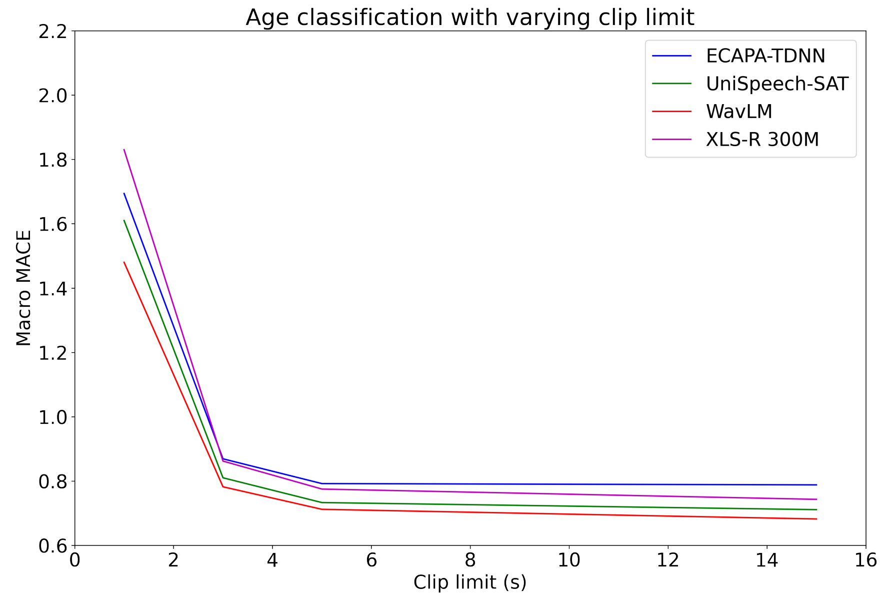
In Section 2.2.1, we discussed the clip length distribution of the CVXL Corpus. We showed that the voice clips have a mean length of 5.2 seconds and a standard deviation of 1.9 seconds. As even the slowest of our models show a throughput of 28-32 samples per second for 10 seconds voice clips, we have a sufficiently low inference time across all data on a GPU. The inference time on a CPU is 20-50 times slower, up towards 1.2 seconds for a 10 second clip. This is quite slow, but could be feasible in certain situations, especially if an earlier part of a conversation can be used for the inference. Thus, our results have a practical value and can predict both age and gender in real time using a GPU and potentially even with a CPU.

4.3.2 Clip length

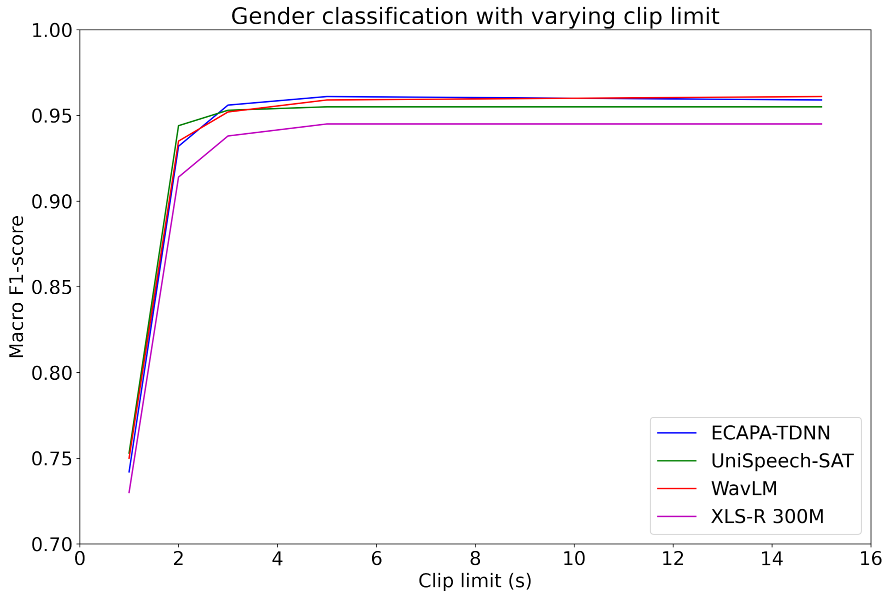
In our research, we have made the assumption of full length voice clips being available for use, with sentences spoken from start to finish. Having fully spoken sentences is not necessarily true in the context of conversational applications. For instance, in the important application of real time analysis, the input is continuously growing. The speaker will therefore be analyzed while speaking and consequently, the clip lengths fed to the model will affect the model's performance.

There will always be a trade-off between time and information as longer clips take longer to analyze, but contain more information about the speaker. Therefore, we can further explore the practicality of our models by investigating the aspect of model performance relative to clip length.

Figure 4.3a shows the macro average MACE scores for our fine-tuned models across different clip limits in the age classification task. Correspondingly, Figure 4.3b shows the macro F1-scores for gender classification. The clip limits denote the initial seconds used of the voice clips and thus the last data-point corresponds to having the original lengths of the voice clips, i.e. capped at 15 seconds. These data-points are equivalent to the previous fine-tuning results in Table 4.4 and Table 4.5 for age and gender classification respectively. The dataset used is the monolingual dataset CVXL English.



(a) Age classification macro average MACE for the fine-tuned models with max clip lengths of 1, 3, 5 and 15 seconds on the CVXL English dataset.



(b) Gender classification macro average F1-score for the fine-tuned models with max clip lengths of 1, 2, 3, 5 and 15 seconds on the CVXL English dataset.

Figure 4.3: Results from the fine-tuned models for different clip lengths. The clip limits denote the initial seconds of the voice clips used.

A clear improvement can be seen as longer clips are being used for both gender and age classification. Results using the first 5 seconds of the clips for age classification, and 3 seconds for gender, are almost as good as using the full length clips. This indicates that longer clips are not essential to performance, as after a certain threshold the performance is no longer limited by clip length.

It is worth noting that there is some uncertainty around what part of the clips contain speech. Often, the clips begin with some silence, which clearly will influence the results when using only the first part of the clips. This partially explains the very poor results achieved when only using the first second, and also means we cannot assume the results in Figure 4.3 reflect the amount of speech needed for a certain performance.

Regardless, this shows that our models are applicable in a real-time setting, as only a small amount of speech is necessary for close to optimal performance.

4.3.3 Age classification with three classes

Another aspect of value for applicability is the number of age groups used. We trained on seven age classes, combining the higher age groups due to the low support, as this is how our data was structured. However, the significance of being able to distinguish between age groups of ten years is questionable. Distributing speakers into young, middle-aged and old age groups could be just as applicable and is a simplification of the problem.

In Table 4.11, we show binned results from our fine-tuned models, where the seven age groups has been divided into young (10-29), middle aged (30-59) and old (60+). To evaluate this three class model, we use a regular macro average F1-score instead of MACE, as an average class error metric has little value with only three classes.

Table 4.11: Macro average F1-scores for three age classes (young (10-29), middle aged (30-59) and old (60+)) when the fine-tuned results from ordinal classification of seven classes were binned into the three classes.

Model	English	1b	5b	20b
ECAPA-TDNN	0.713	0.635	0.686	0.671
UniSpeech-SAT	0.717	0.652	0.731	0.683
WavLM	**0.725**	**0.704**	**0.733**	0.671
XLS-R 300M	0.705	0.700	0.726	**0.711**

The three class model achieves decent results with fairly even performance across both monolingual and multilingual datasets. WavLM seems to outperform the other models slightly, except for on CVXL 20b where XLS-R 300M is better.

By reducing the problem down to three classes instead of seven, we seemingly simplify our problem, as we remove boundaries between our classes, but yet we only see decent results. When grouping classes, we consequently also put more emphasis on the remaining boundaries, as it is now vital for our model to distinguish between the age groups on the edges of the three class division. For example, if our model is poor at distinguishing between people in their twenties and thirties, in a seven class setting this might still yield an acceptable result in terms of class error. In a three class setting an incorrect prediction with one class error has a greater impact on the end result. The positive aspect of applying the models in this way is the extreme rarity of classifying an old person as young and vice versa. All the models are really good at not making this mistake and having a three class problem exemplifies that well.

Chapter 5
Discussion

The field of speech processing is undeniably growing fast. All of the pre-trained models that we have utilized, except for ECAPA-TDNN, did not exist at the start of our research and were released during the course of it. Thus, it's not unreasonable that there will be new models in the near future that could outperform the models tested in this thesis.

To the best of our knowledge, there is no comparable work on age and gender classification with language-agnostic models. This puts us in a position where it is somewhat difficult to judge the quality of our results. In a monolingual setting, we can compare our results to previously published results on the TIMIT dataset, but the multilingual aspect remains unexplored. We rather hope our work will provide a baseline for these uncommon speech processing tasks in this rapidly growing field.

5.1 Pre-trained embeddings

For the gender task, the pre-trained embeddings seem to contain a lot of relevant information as the results are fairly close to optimal. A macro F1-score above 96% is slightly below what has been performed on other datasets in the literature, but higher than some results on notoriously noisy datasets such as aGender. Thus, as the CVXL corpus is notably noisy we believe this is a very good result, and certainly good enough to apply in a conversational application.

Interestingly, for the gender classification task, the ECAPA-TDNN model performs just as well as the heavier transformer models. If one were to only be interested in the gender of the speaker, this would be the way to go given that the inference time and memory requirements are significantly lower. This can probably be attributed to the simpler nature of the gender classification problem, and perhaps one could assume the attributes in voice pertaining to gender are related to the vocal tract and voice quality to a larger extent. This would make sense as the simpler F-bank features used by the ECAPA-TDNN represent such qualities well, but lack ways to represent context and diction.

For the age task, we wanted to try both regression and ordinal classification before deciding what to use later during fine-tuning. When binning the results of the regression to the classes, it became clear that the two approaches performed relatively equally. Thus, ordinal classification was assumed to be superior as it suits the discrete nature of the Common Voice data better.

5.2 Fine-tuned embeddings

We fine-tuned by training the pre-trained models on the age classification task for all CVXL datasets respectively. The training now utilized a task not used before in training of the embedders, which is expected to improve the embedding quality for that specific task.

Unsurprisingly, the model that has the best general and speaker related embeddings according to the SUPERB benchmark, WavLM, generally outperforms all the other models. It is also the newest model, and only became available late into the process of writing this thesis. Further, UniSpeech-SAT performs overall better than both ECAPA-TDNN and Wav2Vec2-XLS-R 300M taken over all the datasets. This can probably be attributed to the fact that ECAPA-TDNN is getting outdated by processing F-bank features rather that the sound signal itself, and that the Wav2Vec2-model is pre-trained on automatic speech recognition (ASR) rather than a speaker related task. Again, for gender recognition, ECAPA-TDNN performs as well as the other models, and could therefore be preferable due to its lower requirements if gender is the only task of importance.

A central question for this thesis is how language independent the resulting models are, motivating the creation of the different CVXL datasets. The performance on the English dataset is meant to give an indication of how well the model can perform on a language it has seen plenty of, whereas the other datasets tell us how well the models generalize to unseen languages. For both age and gender, the models perform significantly better on the English-only dataset, although not much for gender. Training only on English performs the worst on unseen languages, and training on the 5 or 20 biggest languages performs fairly equally.

Much of the variability in the results, for example that the models perform slightly better on CVXL-5b than on CVXL-20b, can be attributed to chance. The difficulty of the validation and test sets unfortunately seems a bit random. Having a single multilingual evaluation set for all the datasets might have yielded clearer results on the performance gain of training on more languages. However, when using the 20b test set to test all models, see Table 4.6, one can assume that including more languages in the training is most likely only positive for performance on unseen languages.

The outstanding results on the TIMIT dataset merely show how fast the embedding models in this field are improving. They are now great at creating universal speech representations, good enough to be utilized in tasks unrelated to the way they were pre-trained. Further, the results are also much better relative to the results on our CVXL datasets. This can be attributed to the cleanliness of the TIMIT dataset, which was produced in a research setting, in contrast to the crowd-sourced nature of Common Voice. It certainly highlights the sensitivity to noisy data of these models, further solidifying the need for augmentation when training them.

Due to time-constraints and the relatively optimal results found using pre-trained embeddings for gender classification, we chose not to move on with fine-tuning the embedders for this task. Although this might have improved the performance on gender classification even more than the age fine-tuning did, these results were deemed good enough. Also, using the same embedder for both tasks allows a single model with two separate heads, optimizing inference time.

This transfer learning approach is likely a better approach than trying to build high performing models from scratch, as most of these models have been pre-trained on large amounts of data, often in a self-supervised manner. The amount of computational resources and time needed to match that is simply too large for this project.

5.3 The dataset

Although the CommonVoice project provided us with an abundance of free-to-use data, this data also came with its problems. First and foremost, only using age bins rather than exact age was a consistent problem for this thesis, as it introduced questions about how to approach this. We are satisfied with our solution, namely ordinal classification, but presumably better results in terms of mean absolute error would have been achieved with exact ages. Second, the crowd-sourced nature of the data is in a way a good thing, given the variability in recording conditions. However, the accuracy of the reported age and gender can be questioned.

Furthermore, the heavy imbalance toward men in their twenties is another aspect that we had to deal with. The main approach we used to combat the imbalance was a simple bins sampling strategy, i.e. sampling all classes equally often during training. This made the models equally good at all the classes, although presumably with less generalization for the smaller, over-sampled classes. However, augmentation proved an effective tool to still make these classes generalize well.

5.4 Future work

Inference time is an important aspect of the practicality and applicable value of our models. In our research, we used large models, up towards 300 million parameters, as these were regarded as the current state of the art. These still had an inference time short enough for usage in real-time settings on a GPU. There exists smaller models, as well as smaller versions of the same models, that would be of interest to fine-tune. Although, we did not evaluate them, these models probably have a lower performance, but with a substantial improvement in inference time. Depending on the loss in performance, these smaller models might have more of an applicable value than the models we tested. Additionally, various pruning methods could potentially be applied to the large models in order to boost inference time while preserving performance.

With their publications, both UniSpeech-SAT (Chen et al., 2021b) and WavLM (Chen et al., 2021a) performed a weight analysis of the networks. They showed that shallow layers seem to contribute more for tasks such as speaker verification and diarization, while top layers are of importance for automatic speech recognition and intent classification. That is, more information about the speaker exists in the shallow layers, while semantic and content information is skewed toward the top layers. Therefore, a weighted average of the layer outputs with learned weights could be implemented in order to use different parts of the network. This could help extract more valuable features for both our tasks, where speaker information is at the core. Keeping the shallow layers frozen during fine-tuning could also be of interest. This would preserve the pre-trained information and only change the higher level semantics to, in theory, be age specific.

5.5 Conclusion

The applicability of our models in voice-based conversational applications was the main focus of this thesis. We wanted to build machine learning models, capable of both accurate predictions and fast inference on relatively short voice clips, preferably even on previously unseen languages. Thus, we evaluated our fine-tuned models with both monolingual and multilingual datasets.

In age classification, across all of our different datasets, our models reached a macro average MACE which is below 1. On average, our models predict an age less than 1 age group away from the ground truth. The error is even smaller in a monolingual setting, with WavLM reaching a 0.682 macro average MACE (see Table 4.4), while the multilingual datasets proved to be more difficult. Even here, XLS-R 300M reaches 0.886 macro average MACE with both UniSpeech-SAT and WavLM not far behind with a macro average MACE of ~0.92. Although not a completely correct comparison, we think this is equivalent to an average error of ~7 years and ~9 years for the monolingual and multilingual datasets respectively, which certainly are applicable results.

For gender classification, as shown in Table 4.5, the macro F1-scores are ~0.96 across all datasets and do not vary much between our different models. This gives us a high certainty that our models are predicting speakers' genders correctly, well within usability performance in conversational applications.

Additionally, we did experiments on inference time and clip lengths in Section 4.3 to further investigate our models' applicability. We showed that inference on a GPU is very fast, with ~30ms on 10 seconds voice clips for the transformer models and even faster at ~9 ms for ECAPA-TDNN. For a CPU, we get up towards 50 times slower inference speed, with all the models requiring 1.2 seconds for a 10 seconds clip.

Regarding clip lengths, we achieved similar results with a 5 second clip limit as with full-length clips for age classification. For gender classification, the performance converged faster and we showed that a 3 second clip limit gave similar macro average F1-scores as with full-length clips. However, due to the uncertainty of which part of the clips contain speech, too much emphasis should not be put on these specific clip lengths. We rather conclude that a threshold likely exists where more speech in each clip does not improve performance significantly, and that this threshold is located at a fairly small amount of recorded speech.

Combining the results of accuracy, inference speed, and clip lengths, we have demonstrated the applicability of our models. With only a few seconds of recorded speech in any given language, we can predict a person's age and gender with a GPU, and potentially even with a CPU, in a short amount of time with a reasonably small error. Thus, we have answered the questions posed in our problem statement, and there seems to be potential for these models in real-time conversational applications.

References

Alkhawaldeh, R. S. (2019). Dgr: Gender recognition of human speech using one-dimensional conventional neural network. *Scientific Programming*, 2019.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint, arXiv:2111.09296.*

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint, arXiv:2006.11477.*

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., and Wei, F. (2021a). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint, arXiv:2110.13900.*

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., and Yu, X. (2021b). Unispeech-sat: Universal speech representation learning with speaker aware pre-training. *arXiv preprint, arXiv:2110.05752.*

Chen, Z., Chen, S., Wu, Y., Qian, Y., Wang, C., Liu, S., Qian, Y., and Zeng, M. (2021c). Large-scale self-supervised speech representation learning for automatic speaker verification. *arXiv preprint, arXiv:2110.05777.*

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech 2020.*

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint, arXiv:2106.07447.*

Kwasny, D. and Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14).

Mishra, M. K. and Shukla, A. K. (2017). A survey paper on gender identification system using speech signal. *International Journal of Engineering and Advanced Technology (IJEAT)*, 6(6).

Nasef, M. M., Sauber, A. M., and Nabil, M. M. (2021). Voice gender recognition under unconstrained environments using self-attention. *Applied Acoustics*, 175:107823.

Okabe, K., Koshinaka, T., and Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. *Interspeech 2018.*

Qawaqneh, Z., Mallouh, A. A., and Barkana, B. D. (2017). Deep neural network framework and transformed mfccs for speaker's age and gender classification. *Knowledge-Based Systems*, 115:5–14.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv preprint, arXiv:2106.04624.*

Shue, Y.-L. and Iseli, M. (2008). The role of voice source measures on automatic gender classification. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4493–4496.

Smith, L. N. and Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In Pham, T., editor, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 369 – 386. International Society for Optics and Photonics, SPIE.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., tik Lee, K., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and yi Lee, H. (2021). Superb: Speech processing universal performance benchmark. *arXiv preprint, arXiv:2105.01051.*

Appendices

Appendix A

Common Voice XL

Table A.1: The amount of training and testing data for all age groups in each CVXL dataset. The total amount of data in each age group for the whole CVXL corpus is also shown. Sizes are represented in minutes.

Age	English	1b	5b	20b	All
10-19	784.6	821.5	1269.8	1631.8	1727
20-29	2907.9	2948.4	4733.3	6979.5	7743.3
30-39	1455.5	1499.5	2778.1	3937.8	4382
40-49	782.1	825.3	1862.7	2383.9	2556
50-59	502.9	546.5	1286.8	1541.6	1604.5
60-69	315.5	361.6	694.7	806.7	806.7
70-79	117.9	169.5	234.8	234.8	234.8
80-89	15.3	27.2	27.2	27.2	27.2
90-99	0.9	4.0	4.0	4.0	4.0
Total	6,883	7,203	12,891	17,547	19,086

Table A.2: The amount of training and testing data for male and female in each CVXL dataset. The total amount of male and female data the CVXL corpus is also shown. Sizes are represented in minutes.

Gender	English	1b	5b	20b	All
Male	5,382	5,571	9,942	13,314	14,410
Female	1,500	1,633	2,949	4,233	4,676
Total	6,883	7,203	12,891	17,547	19,086

Table A.3: Languages used for training and testing together with their respective size in each dataset. The total support in the Common Voice XL corpus is also shown. Sizes are represented in minutes.

Language	**ISO**	**English**	**1b**	**5b**	**20b**	**All**
Abkhaz	ab	-	4.0	4.2	4.5	5.7
Arabic	ar	-	4.0	4.3	5.3	117.2
Armenian	hy-AM	-	2.0	2.1	2.6	4.7
Assamese	as	-	1.4	1.5	1.9	3.9
Bashkir	ba	-	5.2	6.6	11.1	121.7
Basque	eu	-	7.1	8.5	14.2	163.9
Basaa	bas	-	0.3	0.3	0.3	0.3
Belarusian	be	-	8.0	9.0	530.8	530.8
Bulgarian	bg	-	2.6	2.8	3.6	9.0
Breton	br	-	3.2	3.4	4.1	13.2
Catalan	ca	-	14.3	1018.4	1018.4	1018.4
Chinese (China)	zh-CN	-	3.2	3.7	342.5	342.5
Chinese (Hong Kong)	zh-HK	-	5.2	5.5	183.7	183.7
Chinese (Taiwan),	zh-TW	-	3.6	3.9	149.0	149.0
Chuvash	cv	-	3.1	3.2	4.4	10.6
Czech	cs	-	3.7	3.9	5.7	68.1
Dhivehi	dv	-	3.1	3.4	4.5	42.9
Dutch	nl	-	5.9	6.7	200.6	200.6
English	en	6882.7	6882.7	6882.7	6882.7	6882.7
Esperanto	eo	-	9.7	11.0	200.2	200.2
Estonian	et	-	5.8	6.0	237.1	237.1
Finnish	fi	-	3.0	3.4	4.5	14.2
French	fr	-	13.2	1558.7	1558.7	1558.7
Frisian	fy-NL	-	5.1	6.0	17.4	58.4
Galician	gl	-	3.1	3.1	4.8	16.4
Georgian	ka	-	2.5	2.8	3.1	16.5
German	de	-	15.7	1744.2	1744.2	1744.2
Greek	el	-	3.0	3.1	4.6	22.9
Guarani	gn	-	1.2	1.3	1.7	2.3
Hakha Chin	cnh	-	2.5	2.6	3.5	19.2
Hausa	ha	-	1.2	1.2	1.5	2.2
Hindi	hi	-	2.6	2.7	3.5	27.2
Hungarian	hu	-	4.0	4.1	5.1	18.8
Indonesian	id	-	2.9	3.0	4.1	45.8
Interlingua	ia	-	3.4	3.5	4.2	6.6
Irish	ga-IE	-	3.0	2.9	3.9	14.2
Italian	it	-	14.2	18.7	854.3	854.3
Japanese	ja	-	3.1	3.3	5.3	56.3
Kabyle	kab	-	6.9	7.6	10.3	116.9
Kazakh	kk	-	2.3	2.6	2.9	4.1
Language	**ISO**	**English**	**1b**	**5b**	**20b**	**All**

Kinyarwanda	rw	-	3.8	3.7	291.2	291.2
Kurmanji Kurdish	kmr	-	3.7	3.8	5.4	39.1
Kyrgyz	ky	-	2.2	2.1	3.0	33.1
Latvian	lv	-	2.2	1.9	3.2	13.8
Lithuanian	lt	-	4.7	4.9	7.0	42.5
Luganda	lg	-	2.4	2.5	3.2	40.2
Maltese	mt	-	4.7	4.9	6.9	27.5
Mongolian	mn	-	3.0	3.1	4.6	43.0
Odia	or	-	1.9	2.0	2.7	3.7
Persian	fa	-	4.0	4.5	528.2	528.2
Polish	pl	-	4.5	4.8	240.0	240.0
Portuguese	pt	-	5.0	6.0	215.4	215.4
Punjabi	pa-IN	-	2.0	2.1	2.3	3.1
Romanian	ro	-	3.5	3.7	4.8	35.5
Romansh Sursilvan	rm-sursilv	-	3.3	3.2	4.2	8.7
Romansh Vallader	rm-vallader	-	2.7	2.9	3.6	4.3
Russian	ru	-	5.7	5.5	313.5	313.5
Sakha	sah	-	3.8	4.1	5.1	6.9
Serbian	sr	-	1.0	1.0	1.1	1.7
Slovak	sk	-	2.0	2.2	2.8	10.5
Slovenian	sl	-	3.9	4.3	4.9	15.3
Sorbian Upper	hsb	-	3.2	3.4	3.6	5.2
Spanish	es	-	10.1	1396.4	1396.4	1396.4
Swedish	sv-SE	-	3.9	4.2	5.5	51.9
Tamil	ta	-	6.8	7.3	9.5	111.6
Tatar	tt	-	4.7	5.2	6.4	18.3
Thai	th	-	3.7	4.2	241.5	241.5
Turkish	tr	-	3.8	4.1	5.2	114.7
Ukrainian	uk	-	4.8	5.7	7.2	121.4
Urdu	ur	-	1.2	1.3	1.4	3.8
Uyghur	ug	-	4.7	4.6	6.3	25.8
Uzbek	uz	-	0.6	0.7	0.8	1.2
Vietnamese	vi	-	2.7	3.0	3.6	9.2
Welsh	cy	-	11.8	12.5	162.0	162.0

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2022-03-15

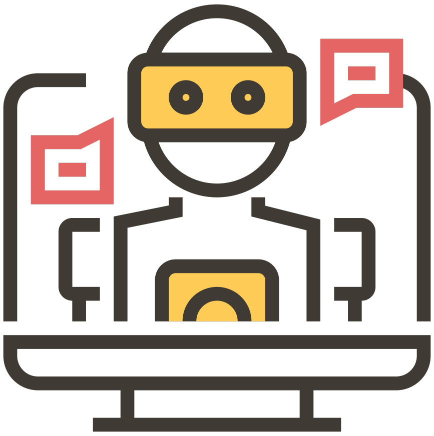
EXAMENSARBETE Language Agnostic Voice Classification for Conversational Applications
STUDENTER Edwin Ekberg, Fredrik Lastow
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

Röstbaserad klassificering av kön och ålder

POPULÄRVETENSKAPLIG SAMMANFATTNING **Edwin Ekberg, Fredrik Lastow**

Människor har en naturlig förmåga att höra vem en talar med, endast genom att höra dennes röst. Specifikt har vi förmågan att höra vilket kön och ungefär vilken ålder personen vi talar med har. Vi ställde oss frågan om maskininlärningsmodeller skulle kunna göra samma sak, och på så sätt hjälpa konversationsbaserade applikationer att anpassa sina svar till användaren.

Som vanligt i ett maskininlärningsprojekt behövs det data. Vi valde att utnyttja det crowd-sourcade och publika datasetet Mozilla Common Voice. Vi filtrerade ut alla röstklipp som hade data på både talarens kön och ålder, vilket resulterade i hela 318 timmar inspelad röst. Den stora fördelen med detta dataset är de diversifierade inspelningsförhållandena som skapar naturlig generalisering, medan en stor nackdel är ett kraftigt bias mot unga män.



Vi valde att testa modeller som har förtränats på liknande uppgifter, exempelvis talaridentifiering och verifiering. Hypotesen var att modeller som på något sätt destillerar information om talaren kan bidra med bra embeddings för köns- och åldersbestämning. Detta resulterade i att vi testade fyra olika modeller med varierande storlek och förträningsprocedurer.

Efter att vi testat prestationen hos de förtränade nätverken, förbättrade vi också resultaten genom att träna vidare nätverken på åldersklassificering. Det visade sig att ännu bättre embeddings (för våra användningsområden) kunde skapas genom att göra detta. Könsbestämmning kunde med hjälp av dessa göras korrekt i ungefär 96% av fallen, medan åldersklassificeringen i snitt gissade mindre än 10 år fel.

För att kunna jämföra våra resultat med tidigare publicerade sådana valde vi även att träna på det flitigt använda TIMIT datasetet. Här slog vi de bästa resultaten i literaturen, vilket föreslår att våra resultat på Common Voice borde vara konkurrenskraftiga.

Modellerna bedömdes fullt applicerbara i realtidsapplikationer då endast en liten mängd tal behövs för nästintill optimal prestanda och processtiden är mycket kort.