B. TECH. PROJECT REPORT

On

Estimation of speaker characteristics from speech signal

By
Tarun Gupta

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
May 2022

Estimation of speaker characteristics from speech signal

A PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degrees

of
BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

Submitted by:
Tarun Gupta

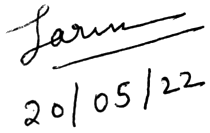
Guided by:
Dr. Ranveer Singh (IIT Indore)
Dr. Chng Eng Siong (NTU Singapore)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
May 2022

CANDIDATE'S DECLARATION

I hereby declare that the project entitled **"Estimation of speaker characteristics from speech signal"** submitted in partial fulfillment for the award of the degree of Bachelor of Technology in 'Computer Science and Engineering' completed under the supervision of **Dr. Ranveer Singh, Assistant Professor, Computer Science and Engineering, IIT Indore** and **Dr. Chng Eng Siong, Associate Professor, School of Computer Science and Engineering, NTU Singapore** is an authentic work.

Further, I declare that I have not submitted this work for the award of any other degree elsewhere.



20/05/22

Tarun Gupta
180001059

CERTIFICATE by BTP Guide

It is certified that the above statement made by the student is correct to the best of my knowledge.

Dr. Ranveer Singh
Assistant Professor
IIT Indore

Preface

This report on "Estimation of speaker characteristics from speech signal" is prepared under the guidance of Dr. Ranveer Singh and Dr. Chng Eng Siong.

Through this thesis, I have attempted to provide a description of our approach and methodology to create a deep learning based model for estimation of speaker characteristics, such as age, height and gender from speech signal. This report aims to explain the model architecture, data augmentation and optimization techniques used for the said problem.

The code of the project has been open sourced for public usage and easy reproducibility of our results.

Tarun Gupta
B.Tech. IV Year
Discipline of Computer Science and Engineering
IIT Indore

Acknowledgements

I wish to thank Dr. Ranveer Singh and Dr. Chng Eng Siong for their kind support, expertise and valuable guidance. They were very encouraging and always motivated me, providing constructive feedback throughout the project duration.

I would also like to sincerely thank Mr. Duc-Tuan Truong (masters student, NTU Singapore) for proof-checking my work and critiquing my research ideas and experiments, Mr. Tran The Anh (PhD candidate, NTU Singapore) and Mr. Hexin Liu (PhD candidate, NTU Singapore) for their constant guidance and support.

The National Research Foundation Singapore has funded this study through its AI Singapore Program (Award Number: AISG-100E-2018-006).

The computational work for this study was done in part using Singapore's National Supercomputing Centre's capabilities.

This project report contains material from the following paper in which I am listed as an author.

Gupta, T., Truong, D.T., Anh, T.T. and Siong, C.E., 2022. Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model. arXiv preprint arXiv:2203.11774.

Tarun Gupta
B.Tech. IV Year
Department of Computer Science and Engineering
IIT Indore

Abstract

Estimating speaker attributes like age and height is a difficult task with several applications in speech forensic analysis and potential applications in speaker verification and speaker adaptation techniques. We present a bi-encoder (Mixture of Experts (MoE) inspired) transformer mixture model for estimating speaker age and height in this project. For the extraction of specific male and female voice characteristic features, we suggest the use of two different transformer encoders, while making use of wav2vec 2.0 as a common feature extraction method. The bi-encoder architecture is chosen due to the significant variances in male and female voice characteristics. This architecture increases the model's generalizability by reducing interference effects during the model training. We conduct our tests using the TIMIT corpus and find that our results on age estimation surpass the present state-of-the-art. For male and female age estimation, we obtain 5.54 years and 6.49 years as root mean squared error (RMSE), respectively. Further research into the relative impact of various phonetic sound kinds for speaker profiling reveals that vowel phonemes are the most distinctive for age estimate.

Keywords: speaker characteristic estimation, age estimation, height estimation, wav2vec 2.0, self-supervised representation learning, mixture of experts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Speaker profiling

Speech is an audio output generated by the exact coordination of many human body components. As a result, it has been speculated that acoustic aspects of speech might communicate knowledge about the speaker's physical attributes. Scientific research has looked at the relationship between voice qualities and a speaker's physical attributes such as age and height, among other physical parameters. The sub-glottal resonance frequencies, length of the vocal tract and formant frequencies are all connected to an individual's height [6, 36]. Speech rate, sound pressure level, fundamental frequency, and other voice characteristics change depending on the speaker's age [36, 32]. The speaker's age-related glottis degeneration affects speech features such as jitter, shimmer [25], and speech harmonics [21].

1.2 Applications of speaker profiling

Systems for automatically profiling speakers might be used in a range of sectors. For example, in a criminal investigation, audio recordings might be used to prove a fake bomb threat or a ransom demand over the phone [35, 33]. Estimating the speaker characteristics such height, physical size, age etc. of speakers in audio evidence

might save time for investigative agencies by reducing the number of suspected persons. Further, estimating a speaker's age and gender using voice data might help marketing efforts target the appropriate gender/age client groups [33]. Furthermore, the speaker profile system might be useful in other speech disciplines such as speaker diarization and speech-based verification.

1.3 Contributions

We study how self-supervised based speech representation, especially wav2vec 2.0 [7], may be used to estimate speaker characteristics, viz. height and age. We perform a comparative analysis of wav2vec 2.0 with other feature representations: filter banks, MFCC and X-vectors [38]. To the best of our knowledge, this is the first study to show that wav2vec 2.0 can be used to estimate speaker age and height. We introduce a novel bi-encoder (Mixture of Experts (MoE) inspired) transformer model for the downstream architecture, which uses wav2vec 2.0 as the common feature extraction method followed by a bi-encoder architecture. Bi-encoder design choice was inspired by the disparities between different gender vocal qualities, such as fundamental and formant frequencies [42, 23]. We employ task-dependent uncertainty [17] to formalise our multi-task loss function. To avoid overfitting, we utilise mixup [45] as a regularisation approach. The suggested technique delivers beats previously obtained results on age estimate results on the TIMIT corpus.

Chapter 2

Literature Review

2.1 Traditional approaches

Feature extraction techniques for automated age and height estimation are widely accessible in the literature. However, most earlier research relied on traditional methods for extracting characteristics from raw voice data. For height and age estimates, authors of [24, 9] previously utilised the Open-Smile toolbox to translate short-term spectral information into other statistics such as percentiles, median, mean and so on. The i-vector [30, 44] was used to turn an inconstant-length utterance into a predetermined length embedding vector in a comparable statistical strategy for age and height prediction. Obtaining spectral characterizations of the speech signal is another embedding technique to speaker profiling. As an example, a speaker's resonance frequencies originating from sub-glottal regions are utilised to estimate height [5]. To capture short-term cepstral properties with varied temporal resolutions, Singh et al. [36] used a bag of words representation. Mel Frequency Cepstral Coefficients (MFCC) [28, 13], cepstral and pitch characteristics [25, 16], and other short-term features are also frequently used as speech feature extraction techniques.

Some of the conventional feature extraction techniques used for speaker profiling have been described in the following subsections.

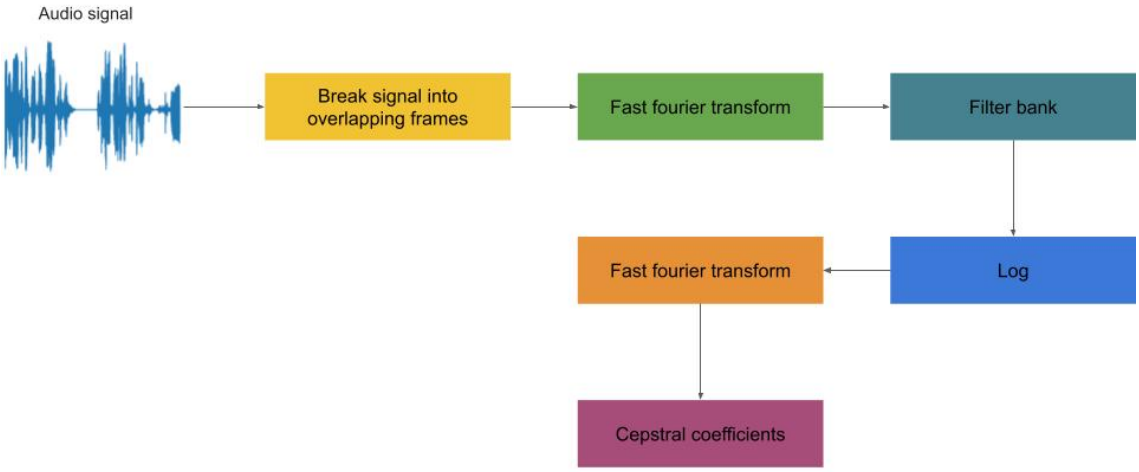


Figure 2.1: Flow diagram illustrating MFCC calculation [2].

2.1.1 Mel-Frequency Cepstral Coefficients (MFCCs)

The Mel-Frequency Cepstral Coefficients (MFCC) feature representation method entails applying pre-emphasis, frame blocking and windowing, applying the Discrete Fourier Transform (DFT) spectrum, taking the log of the output magnitude thus far, and then warping the frequencies using Mel-filter bank, and finally applying the inverse Discrete Cosine Transform (DCT).

In various audio processing models, along with the cepstral coefficients, delta (first order difference) and delta-delta (second order difference) features of MFCCs are also considered as features. The rationale behind using delta and delta-delta features is that it provides dynamics of the power spectrum, that is, the dynamic properties of MFCC over time, which can provide additional information for audio processing tasks. MFCC calculation has been illustrated in Fig. 2.1.

2.1.2 Filter bank

In audio processing, filter bank refers to set of band-pass filters which separate the audio signal into multiple components, with each component carrying a sin-

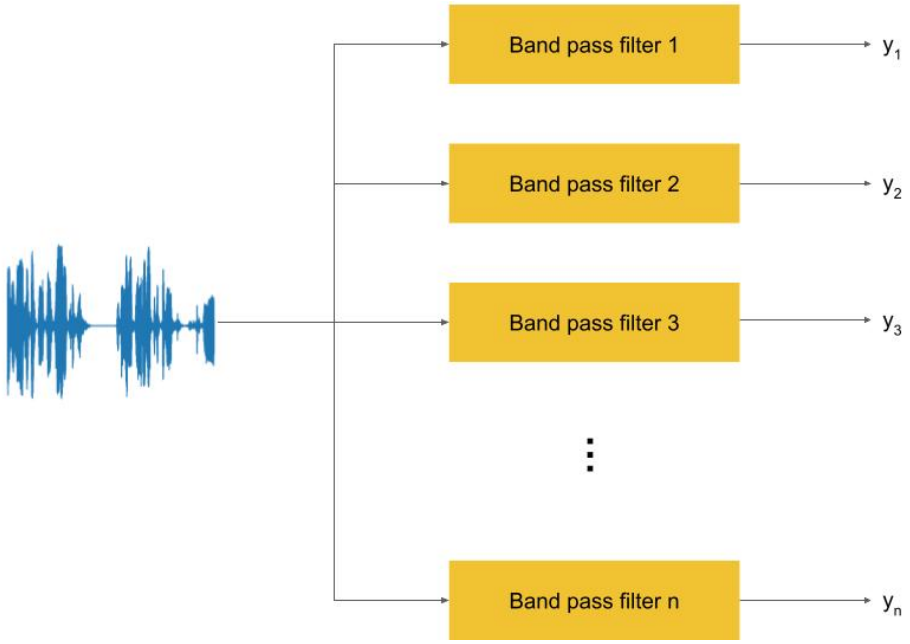


Figure 2.2: Flow diagram illustrating filter bank calculation. A set of band-pass filters with each filter centered at a different frequency separates the audio signal into multiple components.

gle frequency sub-band of the original audio signal [1]. The process of filter bank calculation has been illustrated in Fig. 2.2.

Just as in the case with MFCC, with filter bank as well, often delta and delta-delta features are considered.

2.2 Deep learning based approaches

Deep neural networks (DNN) have demonstrated an exceptional capacity to discover descriptive and unique representations from raw voice audio. As a result, DNN-based speech representation learning has been used in recent research to enhance the performance of the speaker profiling models. Abumallouh et al. [4] demonstrated a speaker gender and age deep-learning based model that, thanks to an unsupervised DNN bottleneck feature extractor, outperforms the original MFCCs feature set, particularly for female speakers. The authors of [19, 20] achieved superior results in age and height estimation in the TIMIT corpus by using DNN discriminative embedding called X-vector [39]. Shangeth et al. [29] employed a semi-supervised

learning paradigm to learn speaker features and produce the best age estimate results on the TIMIT test corpus: 4.8 and 5.0 years as Mean Absolute Error (MAE) for male and female speakers, respectively. A framework for self-supervised learning of speech features and characteristics from raw audio data was recently introduced by Baevski et al. [7]. In various speech domains, such as the realm of speaker recognition, wav2vec 2.0 has significantly improved classification accuracy by capturing far more phonetics information than its predecessor. As per our literature review of speaker profiling research, no work has been done on the usage of wav2vec 2.0 for estimating speaker age and height.

In the following subsections, we briefly describe some of the deep learning based approaches.

2.2.1 X-vectors

X-vectors [38] are deep neural network based discriminative embeddings. The speaker embedding from varied duration utterances is computed using a time-delay neural network (TDNN).

The networks' first five layers operate at the frame level, with a tiny temporal context centered on the current frame t. The size of the final output layer is 512. The statistics pooling layer calculates the mean and standard deviation of all T frame-level outputs from layer frame 5. The statistics are 1500-dimensional vectors that are calculated just once for each input segment. Prior to the non-linearity, X-vectors are retrieved from layer segment 6. [38].

2.2.2 Self-supervised Learning

Self-supervised learning (SSL) is the process of autonomously building a loss from an auxiliary task without the assistance of human annotated labels in order to learn robust features or representations for images, audios and texts. Only the input data is used to build the auxiliary tasks. By optimising the loss function specified by these

auxiliary tasks, deep neural networks may learn resilient features or representations.

In the audio domain, auxiliary tasks include masking the speech input in input or latent space and then solving a contrastive task that requires to contrast between an audio sample from negatives. Examples of this approach include wav2vec [31], wav2vec 2.0 [7] etc.

2.2.3 Self supervised audio representation

We employ self-supervised learning (SSL) based model, viz. wav2vec 2.0 [7], for the speaker profiling tasks. wav2vec 2.0 has shown tremendous success in speech recognition domain, and hence motivated by wav2vec 2.0's success and wide applicability, we study its application in speaker profiling as well. wav2vec 2.0 is able to learn speech features and characteristics in latent space by constructing and solving contrastive challenges. wav2vec 2.0 is made up of a 1D convolution features extraction module $f : X \rightarrow Z$ that takes raw audio input X and generates latent features Z, as well as transformer encoders $e : Z \rightarrow C$ that offer context information. A quantisation module quantizes the feature extractor outputs. Fig. 3.3 shows the architecture of the wav2vec 2.0 model.

2.3 Differences in male and female audio

There are also variances between male and female voices, according to literature. [42, 23] show that the average male fundamental and formant frequencies are lower as compared to female gender voice. As a result, the gender of the speaker influences the extraction of height and age information from speech signals [14]. [20, 29] Most prior speaker profiling research treated gender categorization and height/age estimate as a single challenge. [16, 15] is the only effort to use gender information and gain a minor improvement in speaker profiling task. Nonetheless, these studies just entered the gender value into their model as a binary value. We employ two experts, one for each gender, to capture unique representations of each of the two genders present

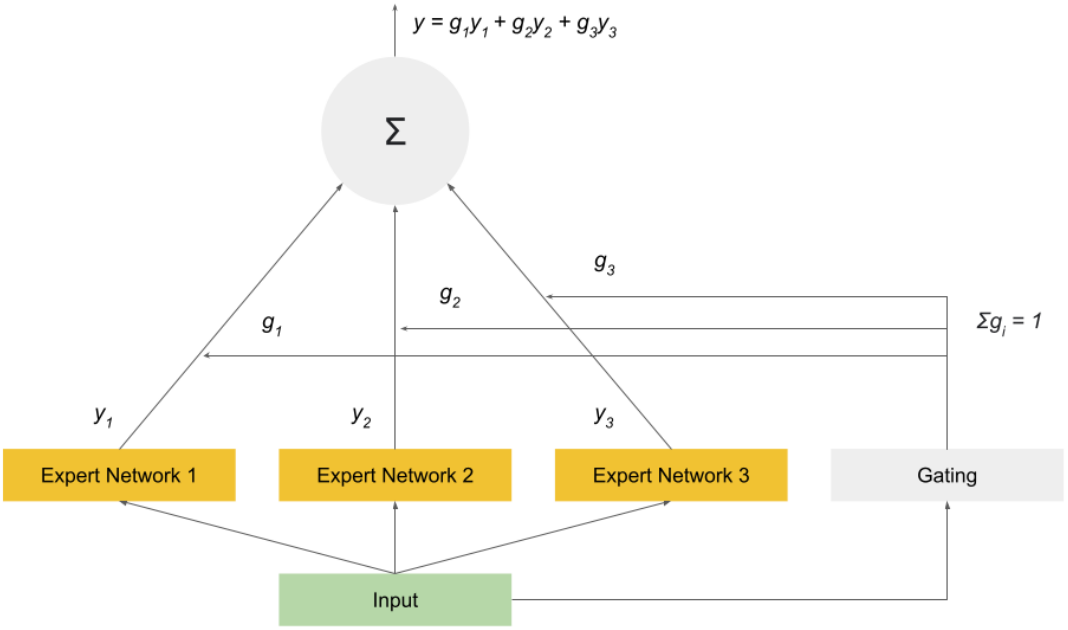


Figure 2.3: Illustration of Mixture of Experts concept

in TIMIT corpus.

2.4 Mixture of Experts

An ensemble learning paradigm where various networks are built to handle different subspaces of the data is known as a mixture of experts (MoE) [11]. The MoE architecture has recently been investigated for a variety of speech tasks, including multi-accent speech recognition [12], code-switching voice recognition [22], and so on.

The divide-and-conquer approach underpins the Mixture of Experts (MoE) adaptive concept [11]. Separate experts of various subspaces can be developed if the training corpus is known beforehand to be naturally partitioned into particular subspaces. The weights to be allocated to each of the expert opinions are determined by a gating network, and then the weighted aggregate of all expert views is calculated. The MoE design reduces backpropagation interference, allowing for quicker training and more generalizability. The MoE architecture is shown in Fig. 2.3.

Chapter 3

Methodology

3.1 Model architecture

In Fig. 3.2, the model architecture is described. First, features are extracted from the raw audio waveform using wav2vec 2.0. On top of the extracted features, a bi-encoder transformer network is developed using the MoE approach. We consider different experts, $MaleExpert$ and $FemaleExpert$, for the two genders included in the TIMIT corpus, female and male. These two experts' architectures are similar, as shown in Fig. 3.2, with six layers of self-attention based transformer encoders and eight attention heads in each layer [40]. To acquire utterance level representations, we perform statistical pooling of the transformer encoder output along the frame dimension. To complete the expert architecture, these utterance level representations are supplied to completely connected layers. The two encoders, i.e., the two experts may concentrate on audio characteristics specific to each gender that are relevant for assessing age and height.

Considering the considerable diversity in auditory characteristics across genders, distinct models for male and female have been created in earlier research [36, 15]. This, however, necessitates the training of two different models, which has the drawback of only using audio samples from one gender at a time. We employ bi-encoder architecture to get the best of both worlds: we can use the entire dataset while still

having distinct experts for male and female gender.

$MaleExpert$ and $FemaleExpert$ supply two expert perspectives, $expert\_view_m$ and $expert\_view_f$:

$$expert\_view_m = MaleExpert(x) \tag{3.1}$$

$$expert\_view_f = FemaleExpert(x) \tag{3.2}$$

$expert\_view_m$ and $expert\_view_f$ are concatenated and given to a completely connected layer, which has a sigmoid activation function to output the gender prediction $g \in [0, 1]$, where x is the representation recovered from wav2vec 2.0. The male gender is mapped to 0 and the female gender is mapped to 1. Prediction of gender serves as a gating network, allowing the two expert's output to be combined in the following way:

$$expert\_view = (1 - g) \times expert\_view_m + g \times expert\_view_f \tag{3.3}$$

To do height and age regression, $expert\_view$ is supplied to completely connected layers.

3.2 Loss function

Three losses are considered into the training process in our multi-task model: height, age, and gender. Prior methods to developing a multi-task architecture for this task used a naive approach of taking a weighted linear summation of these losses [29, 15], with the loss weights fine-tuned manually. To avoid manual fine-tuning of coefficient values of the individual loss functions, we utilise the uncertainty loss [17], which combines several losses using homoscedastic uncertainty. Using this, our loss function is formulated as follows.

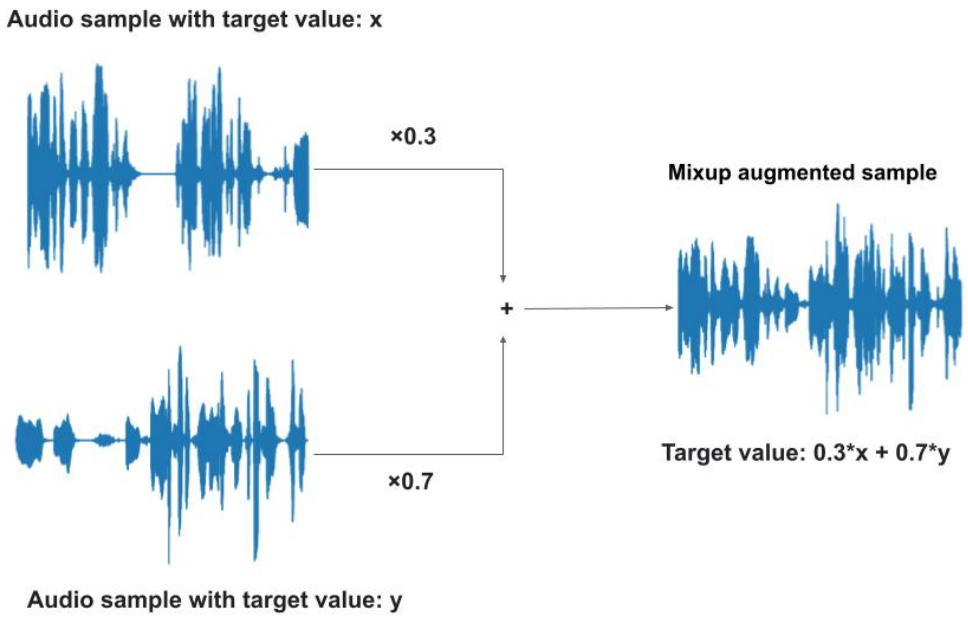


Figure 3.1: Illustration of mixup augmentation.

$$\mathcal{L} = \frac{\mathcal{L}_{height}}{2\sigma_{height}^2} + \frac{\mathcal{L}_{age}}{2\sigma_{age}^2} + \frac{\mathcal{L}_{gender}}{2\sigma_{gender}^2} + log(\sigma_{height}\sigma_{age}\sigma_{gender}) \tag{3.4}$$

where σ_{height}, σ_{age} and σ_{gender} are parameters to be learned. We perform the substitutions $s_{height} = log(\sigma_{height}^2)$, $s_{age} = log(\sigma_{age}^2)$ and $s_{gender} = log(\sigma_{gender}^2)$, as advised in their original work, in the above equation for numerical stability.

3.3 Mixup

We employ mixup as a regularisation strategy. Mixup has been previously been used in various tasks in speech domain such as speaker verification [46].

The mixup augmented sample for two audio samples x_i and x_j, with their corresponding height values as h_i, h_j, age values a_i, a_j, and gender values g_i, g_j, is formulated as follows.

$$x_{mixup} = \lambda x_i + (1 - \lambda)x_j \tag{3.5}$$

$$h_{mixup} = \lambda h_i + (1 - \lambda)h_j \tag{3.6}$$

$$a_{mixup} = \lambda a_i + (1 - \lambda) a_j \tag{3.7}$$

$$g_{mixup} = \lambda g_i + (1 - \lambda) g_j \tag{3.8}$$

here $\lambda \sim U(0, 1)$. The shorter audio is repeated to match it's length with the longer audio to cope with audios of varying durations.

This mixup augmentation has been illustrated in Fig. 3.1. Suppose we have two audio signals with target values x and y and suppose the mixing parameter λ is chosen as 0.3. Then the first audio signal is multiplied by 0.3, and the second audio signal is multiplied by $1 - 0.3 = 0.7$ and added together to get mixup augmented sample. The target value for this augmented sample is defined as $0.3 * x + 0.7 * y$. This augmented sample with its target value is then used during training of the machine learning model.

3.4 Data augmentations

Data augmentations are a set of techniques to artificially increase the training set size by generating new data samples from the existing data samples. Data augmentation techniques are particularly helpful when the training set size is small, as is it is in the case of TIMIT corpus [10], which we use for all our experiments in this project.

The different data augmentation techniques implemented in this project have been described below:

3.4.1 Time stretch

This augmentation changes the speed or duration of an audio signal without affecting its pitch. Time stretch is applied to an audio signal with 50% probability while training, and the stretch range of the audio signal is randomly chosen between 0.8 and 1.25.

3.4.2 Pitch shift

This augmentation raises or lowers the original pitch of the audio signal. Pitch shift is applied to an audio signal with 50% probability while training, and the shift range of the audio signal is randomly chosen between positive or negative 4 semitones.

3.4.3 Adding Gaussian noise

This augmentation adds Gaussian noise to audio signal with 50% probability while training. The amplitude of the added noise is chosen randomly from the range 0.001 and 0.015.

3.4.4 Time masking

This augmentation makes a randomly chosen part of the audio silent. This augmentation is applied with 50% probability to an audio signal, and the fraction of audio signal to be masked is chosen randomly between 0.0 and 0.5.

3.4.5 Frequency masking

This augmentation masks some frequency bands on the spectrogram. It is applied with 50% probability to an audio signal. This augmentation essentially applies a band pass filter, with the fraction of bandwidth to be masked is chosen randomly between 0.0 and 0.5.

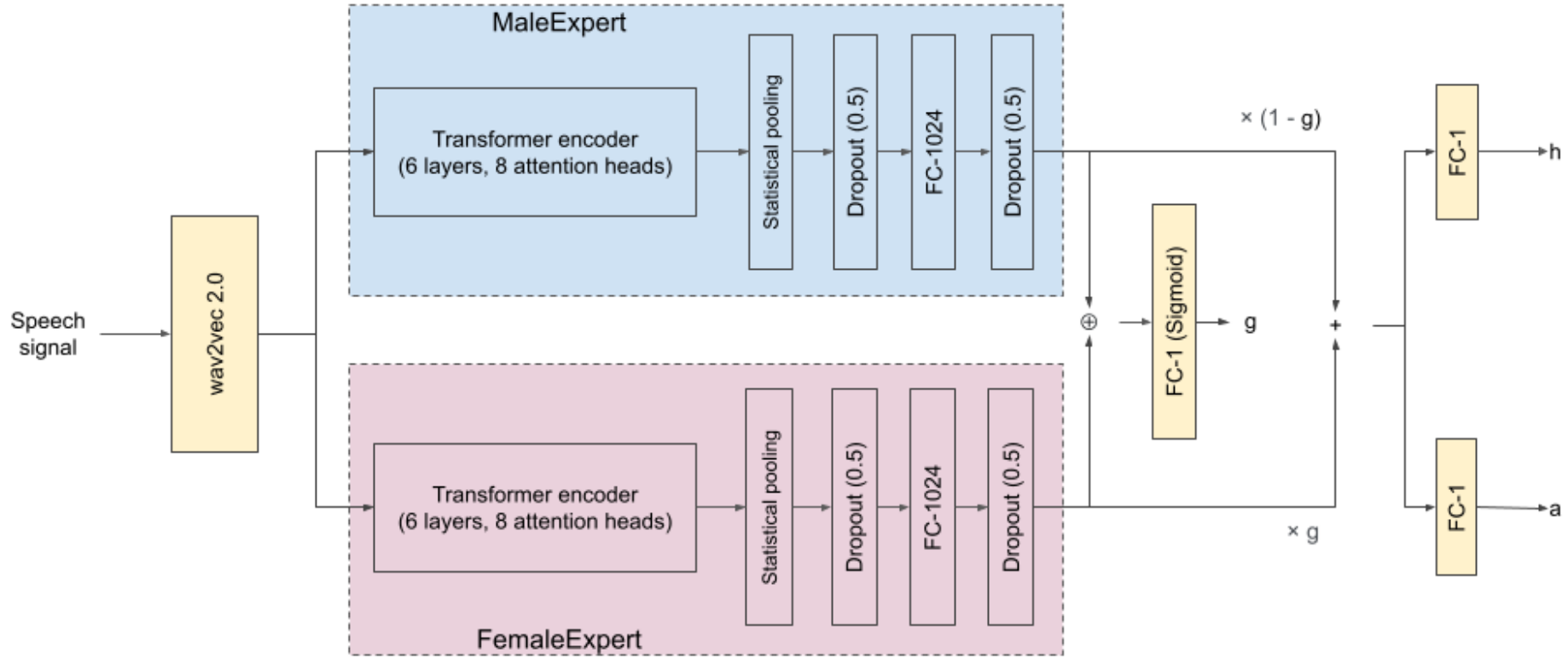


Figure 3.2: Model architecture for wav2vec 2.0 bi-encoder. Self-attention based transformer encoder [40] with 6 layers and 8 attention heads is employed. Pooling here refers to the process of concatenating mean and standard deviation across frame dimensions to create utterance level representation. A completely connected layer having L neurons is designated as '$FC - L$'. The concatenation operation is denoted by $\oplus$. The age, height, and gender predictions are denoted by the letters a, h, and g.

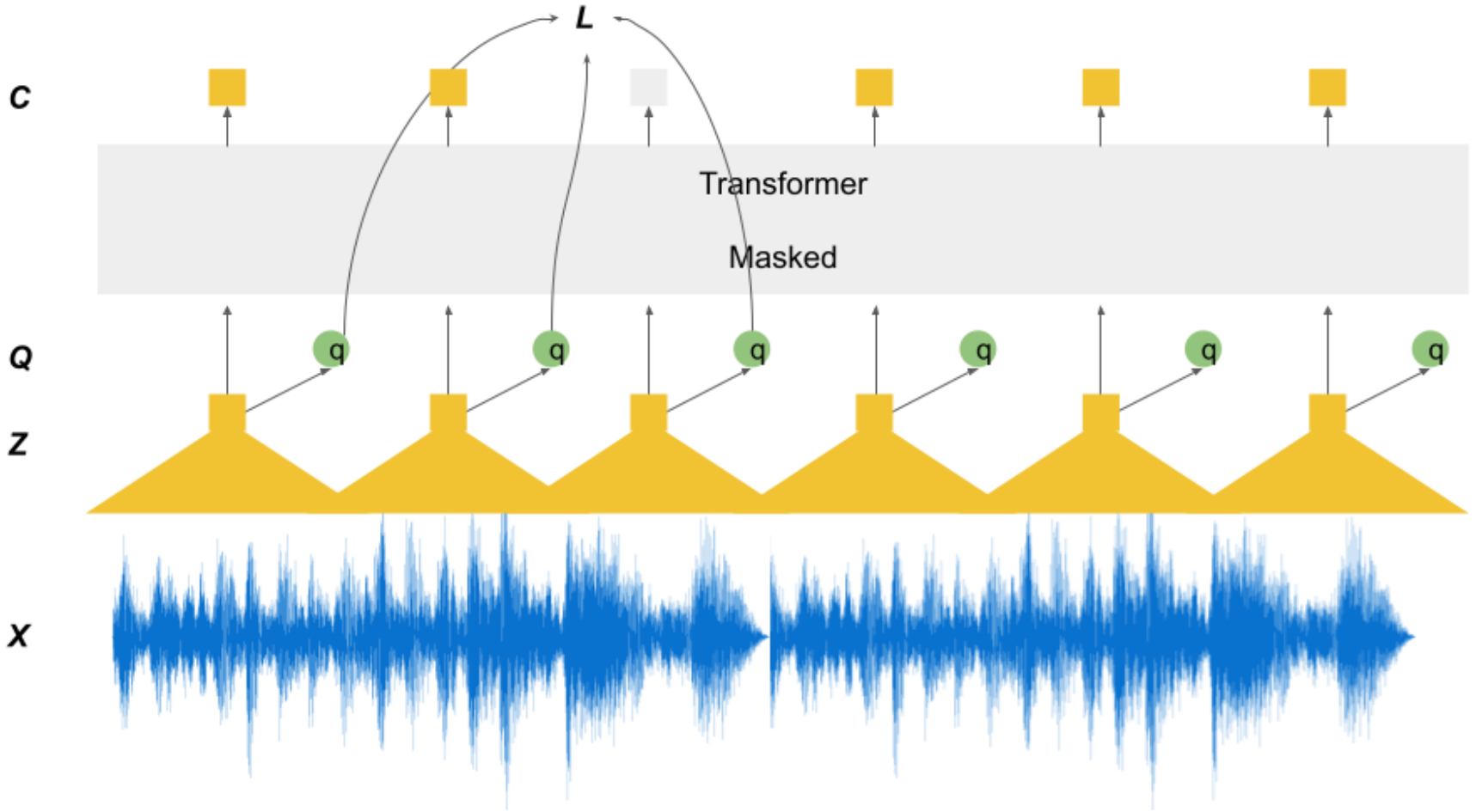


Figure 3.3: The wav2vec 2.0 architecture [7]. Here X represents raw audio waveform, Z represents latent speech representations, Q represents quantised representations and C represents context representations.

Chapter 4

Results and Discussion

4.1 Data corpus

For our investigations, we used the TIMIT dataset [10]. It includes audio clips from 630 people who speak eight distinct American English dialects. The age values vary in from 21 to 76 years in train set, whereas the test set age range is 22 to 68 years. The training set's height range is 145 cm to 199 cm, while the test set's height range is 153 cm to 204 cm. The speech recordings in the collection have an average duration of 2.5 seconds. The TIMIT corpus is pre-divided into evaluation and train sets, making sure there's no common speaker in train and test set.

4.2 Experiment design

Apart from wav2vec 2.0, other characteristics for comparison in the proposed bi-encoder transformer model include filter bank, X-vectors and MFCC. We evaluate 80 mel bins for the filter bank, as well as delta (first order) and delta-delta (second order) features. We use 16 cepstral coefficients, as well as first and second order delta features, in MFCC. Further, Cepstral Mean and Variance Normalization (CMVN) is employed for both the filter bank and the MFCC, with the length of the frame being 25ms and shifting frame by a value of 10ms. We collect frame-level features

for X-vectors before the pooling operation. This is given to the MoE inspired bi-encoder architecture downstream. VoxCeleb1 [26] and VoxCeleb2 [8] training data were used to pre-train the X-vectors.

Next, to demonstrate the effectiveness of MoE inspired bi-encoder architecture, we perform a comparison between the bi-encoder architecture against the single-encoder architecture, with wav2vec 2.0 as the common feature extractor. In wav2vec 2.0 single-encoder model has a single encoder for both male and female audio. The single expert's output is used for estimating speaker characteristics.

In the case of wav2vec 2.0, apart from the initial five convolutional layers of the convolutional feature extraction module, we unfreeze the whole wav2vec 2.0 and use Adam optimizer [18] with learning rate 10^{-6} for models using wav2vec 2.0. We utilise Adam with a learning rate of 10^{-5} when employing MFCC, filter bank, and X-vectors. Mixup technique is typically used for deep learning models and hence should not be considered at the audio level for classic feature extraction techniques such as filter bank and MFCC. Hence, we exclusively employ mixup with wav2vec 2.0 and X-vectors only and not with MFCC and filter bank.

4.3 Results

4.3.1 Comparison of the proposed model with previous works

The findings of the wav2vec 2.0 bi-encoder model are compared to prior efforts in Table 4.1. We perform an extensive comparison of the proposed multi-task model to all of them, whereas many earlier research created distinct models for height and age, or different models for each gender, viz. female and male. Our findings are presented as root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The proposed model gains statistically significant improvement in the age estimate job, as can be observed from Table 4.1. We attain RMSE errors of 5.54 and 6.49 years, respectively, corresponding to an 18.5 percent increase in male age estimate and 8.6 percent improvement in female age estimation over the present state-of-the-

art.

4.3.2 Efficacy of bi-encoder architecture design choice

We wish to illustrate the efficacy of the proposed bi-encoder architecture. To this end, we show the comparison of wav2vec 2.0 (MoE inspired) bi-encoder versus wav2vec 2.0 single encoder models in Table 4.2. In terms of RMSE error, the bi-encoder model outperforms the single-encoder model by 7.0 percent for female age estimation and 2.9 percent for male age estimation, proving our hypothesis of employing distinct encoders for the two genders. However, a similar trend does not appear in height estimate task findings, implying that the MoE inspired bi-encoder design is not effective for height estimate task.

4.3.3 Efficacy of self-supervised representation for speaker profiling

In order to show the wav2vec 2.0 (self-supervised speech representation frameworks) for speaker profiling, we tabulate the findings of several feature extractors in Table 4.4: wav2vec 2.0, MFCC, filter bank, and X-vectors. We can notice that wav2vec 2.0 provides statistically significant improvements as compared to other feature extraction approaches for speaker profiling, as can be shown.

4.3.4 Effect of different augmentation techniques

In Table 4.3, we tabulate the results of the proposed model using five different augmentation techniques: time stretch, pitch shift, adding Gaussian noise, time masking, and frequency masking. The details of their implementation have been discussed in section 3.4. From table 4.3, it can be observed that in all the cases, there's either no or minimal improvement. It can be concluded that the augmentation techniques discussed in this project provide no statistically significant improvement.

4.3.5 Wide band and narrow band comparison

Wide band refers to audio samples with a sampling rate of 16kHz and narrow band refers to audio samples with a sampling rate of 8kHz. TIMIT dataset's audio samples are wide-band. However, most of the telephony audios are narrow-band. Hence, to ascertain the performance of the model on telephony conversations, it's important to test the model on narrow band audios.

To test the model on narrow band audios, we simply down-sample the entire TIMIT dataset to 8kHz and then perform the training and testing. The results are tabulated in Table 4.5. It can be observed that down sampling the TIMIT-corpus to narrow band range doesn't affect the model performance significantly.

4.4 Analysis of phonological significance

To comprehend the significance of different phoneme categories in speaker profiling, we study the performance of the proposed model after masking utterances of a certain phoneme type in the test set of the TIMIT corpus. For each audio sample, the TIMIT corpus provides frame-wise phonetic information. The different types of phones found in the TIMIT corpus are tabulated in Table 4.6. In TIMIT's test corpus, we mask all phonemes in audio samples for each of these phoneme categories, then calculate the age and height RMSE scores. Table 4.6 tabulates the impact of phoneme masking by calculating percentage change that phoneme masking brings, as compared to no masking in the TIMIT test corpus. Due to 'Vowel' masking, we see the highest rise in age RMSE value, indicating that vowel phonemes contain the most knowledge important for age estimate. Perhaps surprisingly, there is no discernible difference in height estimate results, suggesting that height estimate task is independent of phoneme type.

Table 4.1: Comparison of the proposed model with previous works.

Model	**Age RMSE**		**Age MAE**		**Height RMSE**		**Height MAE**	
	Male	Female	Male	Female	Male	Female	Male	Female
Singh et al. [36]	7.8	8.9	5.5	6.5	**6.7**	**6.1**	**5.0**	**5.0**
Kalluri et al. [13]	7.60	8.63	-	-	6.85	6.29	-	-
Kwasny et al. [19]	7.24	8.12	5.12	5.29	-	-	-	-
Williams et al. [41]	-	-	-	-	-	-	5.37	5.49
Mporas et al. [24]	-	-	-	-	6.8	6.3	5.3	5.1
Shangeth et al. (single-task model) [29]	6.96	7.6	4.8	5.1	8.1	6.0	5.9	4.9
Shangeth et al. (multi-task model) [29]	6.8	7.4	4.8	5.0	7.5	6.5	5.8	5.1
Manav et al. (single-task model) [15]	7.20	7.10	5.04	5.02	6.92	6.24	5.20	4.95
Manav et al. (multi-task model) [15]	7.81	8.60	5.50	5.89	6.95	6.44	5.26	5.15
wav2vec 2.0 bi-encoder (ours)	**5.54**	**6.49**	**3.96**	**4.48**	7.3	6.43	5.58	5.07

Table 4.2: Efficacy of bi-encoder architecture design choice.

Model	**Age RMSE**		**Age MAE**		**Height RMSE**		**Height MAE**	
	Male	Female	Male	Female	Male	Female	Male	Female
wav2vec 2.0 single-encoder	5.71	6.98	4.05	4.90	**7.17**	**6.39**	**5.35**	5.08
wav2vec 2.0 bi-encoder	**5.54**	**6.49**	**3.96**	**4.48**	7.3	6.43	5.58	**5.07**

Table 4.3: Effect of different augmentation techniques on wav2vec 2.0 bi-encoder model.

Augmentation	**Height RMSE**		**Height MAE**		**Age RMSE**		**Age MAE**	
	Male	Female	Male	Female	Male	Female	Male	Female
Time Stretch	7.41	6.28	5.69	5.01	5.56	6.49	3.99	4.54
Pitch Shift	7.43	6.49	5.76	5.11	5.66	6.53	4.07	4.87
Gaussian Noise	7.26	6.51	5.48	5.16	5.7	6.79	4.02	4.73
Time Mask	7.33	6.43	5.6	5.02	6	6.85	4.12	4.86
Frequency Mask	7.31	6.3	5.59	4.95	5.75	7.02	4.02	4.86

Table 4.4: Efficacy of self-supervised representation for speaker profiling.

Model	**Age RMSE**		**Age MAE**		**Height RMSE**		**Height MAE**	
	Male	Female	Male	Female	Male	Female	Male	Female
X-vectors bi-encoder	7.66	8.89	5.5	5.82	8.02	6.79	6.11	5.46
filter bank bi-encoder	8.51	8.42	6.19	5.86	7.86	6.68	6.13	5.36
MFCC bi-encoder	8.15	8.65	5.86	6.02	7.63	6.69	5.79	5.33
wav2vec 2.0 bi-encoder	**5.54**	**6.49**	**3.96**	**4.48**	**7.3**	**6.43**	**5.58**	**5.07**

Table 4.5: Wide band and narrow band results comparison.

Model	**Age RMSE**		**Age MAE**		**Height RMSE**		**Height MAE**	
	Male	Female	Male	Female	Male	Female	Male	Female
wav2vec 2.0 bi-encoder wide band	5.58	6.7	4.03	4.66	7.23	6.34	5.55	5.01
wav2vec 2.0 bi-encoder narrow band	5.47	7.31	3.96	4.95	7.28	6.6	5.56	5.24

Table 4.6: Significance of different phoneme types.

Phoneme type	**Age RMSE**		**Height RMSE**	
	Male	Female	Male	Female
Semivowels	12.2%	-0.68%	0.45%	-0.32%
Stops	4.14%	3.05%	0.6%	2.9%
Fricatives	6.08%	-2.41%	1.28%	2.87%
Affricates	0.0%	0.0%	0.0%	0.0%
Others	5.84%	12.17%	1.07%	-2.9%
Nasals	2.51%	-0.27%	-0.52%	0.08%
Vowels	**38.9%**	**20.46%**	2.04%	0.07%

Chapter 5

Implementation

In this section, the implementation of code for this project has been discussed.

5.1 Libraries

5.1.1 PyTorch

PyTorch is an open source library built on top of the Torch library [3]. PyTorch provides an easy and intuitive interface to build deep learning models. PyTorch is developed by Facebook's AI Research lab (FAIR). PyTorch allows us to build deep learning models with a tape-based automatic differentiation system [27].

5.1.2 PyTorch Lightning

PyTorch Lightning is an open-source library that is built on top of PyTorch. It provides a high-level interface for PyTorch, allowing us to spend less time writing boiler-plate code and focusing more of our time on research. It allows us to create scalable deep learning models and run them easily on distributed hardware.

PyTorch in itself is highly flexible and provides an easy interface to build complex

deep learning models. But once the research gets complicated and we have to do distributed optimization, multi-GPU training, etc., it is prone to introduce bugs in the code. PyTorch Lightning solves this exact problem, allowing the user to focus more on research and less on engineering aspects of optimization.

As a result of the above advantages, in this project, we use PyTorch Lightning to implement our models.

5.1.3 S3PRL

S3PRL [43] stands for Self-Supervised Speech Pre-training and Representation Learning. It is an open source toolkit providing various methods for speech representation learning and audio signal processing.

It provides a unified input-output interface for various speech pre-trained models. That is, all the models take input in the same format and provide output in the same format. All the preprocessing specific to a particular model is taken care of by the library.

In this project, we use S3PRL to make use of self-supervised pre-trained models such as wav2vec 2.0 for our model.

5.2 Training

In this section, we describe some of the training choices made in this project.

5.2.1 Multi-GPU training

PyTorch Lightning provides various methods for model training. In this project, we make use of Distributed Data Parallel (DDP) method of multi-GPU training to train our models. In DDP, each GPU initiates its own process, and each of these processes gets access to only a particular subset of the dataset. Each of these

processes initialize the model and perform forward and backward passes in parallel. Then the gradient calculated across these nodes is averaged, which is then finally used to update the optimizer state in each of these nodes.

5.2.2 Learning rate finder

For training deep neural networks, perhaps the most important hyper-parameter that one needs to tune is the learning rate. In order to reduce the guess work in choosing a good initial learning rate, a learning rate finder can be used. Smith [37] described a method to estimate a good learning rate: a small run is done in which the learning rate is increased after each batch, and the corresponding loss is logged. From this, we plot the learning-rate vs loss plot, which can be used to get a good initial learning rate. It's recommended not to choose the learning rate which obtains the lowest loss but instead to choose a learning rate somewhere in the middle of the steepest downward slope.

5.3 Reproducing results

The code for this project has been open-sourced and is available at *github/tarun360/SpeakerProfiling*. For reproducing the results, one can clone this repository and follow the instructions below:

1. Use package manager *pip* to download the requirements using the following command:

   ```
   pip install -r requirements.txt
   ```

2. Download the TIMIT dataset

   ```
   wget https://data.deepai.org/timit.zip
   unzip timit.zip -d 'path to timit data folder'
   ```

```
{
    "model_name": "Wav2vec2BiEncoder",
    "dataDir": {
        "dir": "/notebooks/SpeakerProfiling",
        "data_path": "$dir/TIMIT_Dataset/wav_data/",
        "speaker_csv_path": "$dir/Dataset/data_info_height_age.csv"
    },
    "model_parameters": {
        "batch_size": "8",
        "epochs": "50",
        "lr": "1e-6",
        "model_type": "Wav2vec2BiEncoder",
        "upstream_model": "wav2vec2",
        "_comment": "upstream model to be loaded from s3prl. Some of the upstream models are: wav2vec2, hubert, TERA, mockingjay",
        "feature_dim": "768",
        "narrow_band": false
    },
    "gpu": "-1",
    "n_workers": "0"
}
```


Figure 5.1: A snapshot of the *config.json* file

3. Prepare TIMIT dataset (divide it into train, val, and test sets)

```
python TIMIT/prepare_timit_data.py --path='path to timit data folder'
```

4. Update the *config.json* file to update the upstream model, batch size, GPUs, learning rate, etc. and change the preferred logger in *train.py* files. A snapshot of *config.json* file has been shown in Fig. 5.1.

5. To train your own model on TIMIT dataset for speaker profiling:

```
python train_timit.py --data_path='path to final data folder'
--speaker_csv_path='SpeakerProfiling/Dataset/data_info_height_age.csv'
```

6. To test the trained model:

```
python test_timit.py --data_path='path to final data folder'
--model_checkpoint='path to saved model checkpoint'
```

5.4 Pre-trained model weights

We have uploaded the weights of our pre-trained model on this *Dropbox link*.

To use it, simply download it and run the test script with the *model_checkpoint* path pointing to the downloaded model weights.

Chapter 6

Conclusion and Future work

6.1 Conclusion

We described a bi-encoder transformer model based on a Mixture of Experts (MoE) that employs self-supervised representation, viz. wav2vec 2.0, to perform speaker profiling. The results show that having distinct experts for male and female voices helps decrease interference throughout the training process and produce cutting-edge age estimate results. To integrate numerous losses in our multi-task model, we used the homoscedastic uncertainty principle. We intend to investigate alternative feature extractors and self-supervised learning in the future to increase the accuracy of age and height estimates.

6.2 Future work

One interesting direction to explore in the future is to identify applications of this model in speaker adaptation techniques such as Vocal Tract Length Normalization (VTLN). Such speaker adaptation techniques can be quite useful for various speech tasks, such as automatic speech recognition. [34]

Another direction worth exploring is to try different audio feature representations

such as Phone Posteriorgram (PPG) features etc.

Bibliography

[1] Filter bank. https://en.wikipedia.org/wiki/Filter_bank.

[2] The dummy's guide to MFCC. https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd.

[3] Torch. http://torch.ch/.

[4] Arafat Abumallouh, Zakariya Qawaqneh, and Buket Barkana. New transformed features generated by deep bottleneck extractor and a gmm–ubm classifier for speaker age and gender classification. *Neural Computing and Applications*, 30, 10 2018.

[5] Harish Arsikere, Gary K. F. Leung, Steven M. Lulich, and Abeer Alwan. Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation. *Speech Commun.*, 55:51–70, 2013.

[6] Harish Arsikere, Steven M. Lulich, and Abeer Alwan. Estimating speaker height and subglottal resonances using mfccs and gmms. *IEEE Signal Processing Letters*, 21(2):159–162, 2014.

[7] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Interspeech 2018*, Sep 2018.

[9] Todor Ganchev, Iosif Mporas, and Nikos Fakotakis. Audio features selection for automatic height estimation from speech. In *Hellenic Conference on Artificial Intelligence*, pages 81–90. Springer, 2010.

[10] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

[11] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[12] Abhinav Jain, Vishwanath P Singh, and Shakti P Rath. A multi-accent acoustic model using mixture of experts for speech recognition. In *INTERSPEECH*, pages 779–783, 2019.

[13] Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. A deep neural network based end to end model for joint height and age estimation from short duration speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6580–6584. IEEE, 2019.

[14] Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28, 2020.

[15] Manav Kaushik, Tran The Anh, Eng Siong Chng, et al. End-to-end speaker age and height estimation using attention mechanism and triplet loss. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–8. IEEE, 2021.

[16] Manav Kaushik, Van Tung Pham, and Eng Siong Chng. End-to-end speaker height and age estimation using attention mechanism with lstm-rnn. *arXiv preprint arXiv:2101.05056*, 2021.

[17] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 7482–7491, 2018.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Damian Kwasny and Daria Hemmerling. Joint gender and age estimation based on speech signals using x-vectors and transfer learning. *arXiv preprint arXiv:2012.01551*, 2020.

[20] Damian Kwaśny and Daria Hemmerling. Gender and age estimation methods based on speech using deep neural networks. *Sensors (Basel, Switzerland)*, 21, 2021.

[21] Ming Li, Kyu Han, and Shrikanth Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer, Speech, and Language*, 27, 11 2012.

[22] Yizhou Lu, Mingkun Huang, Hao Li, Jiaqi Guo, and Yanmin Qian. Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts. In *INTERSPEECH*, pages 4766–4770, 2020.

[23] Elvira Mendoza, Nieves Valencia, Juana Muñoz, and Humberto Trujillo. Differences in voice quality between men and women: Use of the long-term average spectrum (ltas). *Journal of Voice*, 10(1):59–66, 1996.

[24] Iosif Mporas and Todor Ganchev. Estimation of unknown speaker's height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.

[25] Christian Müller and Felix Burkhardt. Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *Interspeech 2017*, Aug 2017.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[28] B.L. Pellom and J.H.L. Hansen. Voice analysis in adverse conditions: the centennial olympic park bombing 911 call. In *Proceedings of 40th Midwest Symposium on Circuits and Systems. Dedicated to the Memory of Professor Mac Van Valkenburg*, volume 2, pages 873–876 vol.2, 1997.

[29] Shangeth Rajaa, Pham Van Tung, and Chng Eng Siong. Learning speaker representation with semi-supervised learning approach for speaker profiling. *arXiv preprint arXiv:2110.13653*, 2021.

[30] Seyed Omid Sadjadi, Sriram Ganapathy, and Jason W. Pelecanos. Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5040–5044, 2016.

[31] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[32] Susanne Schötz. Acoustic analysis of adult speaker age. In *Speaker classification I*, pages 88–107. Springer, 2007.

[33] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A. Müller, and Shrikanth S. Narayanan. Paralinguistics in speech and language - state-of-the-art and the challenge. *Comput. Speech Lang.*, 27:4–39, 2013.

[34] Koichi Shinoda. Speaker adaptation techniques for automatic speech recognition. *Proc. APSIPA ASC*, 2011, 2011.

[35] Rita Singh, Joseph Keshet, and Eduard Hovy. Profiling hoax callers. In *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, pages 1–6, 2016.

[36] Rita Singh, Bhiksha Raj, and James Baker. Short-term analysis for estimating physical parameters of speakers. In *2016 4th International Conference on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2016.

[37] Leslie N Smith. Cyclical learning rates for training neural networks. arxiv. *Preprint at https://arxiv. org/abs/1506.01186*, 2015.

[38] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[39] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Keri A Williams and John HL Hansen. Speaker height estimation combining gmm and linear regression subsystems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7552–7556. IEEE, 2013.

[42] Ke Wu and Donald G Childers. Gender recognition from speech. part i: Coarse analysis. *The journal of the Acoustical society of America*, 90(4):1828–1840, 1991.

[43] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[44] Ruben Zazo, Phani Sankar Nidadavolu, Nanxin Chen, Joaquin Gonzalez-Rodriguez, and Najim Dehak. Age estimation in short speech utterances based on lstm recurrent neural networks. *IEEE Access*, 6:22524–22530, 2018.

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[46] Yingke Zhu, Tom Ko, and Brian Mak. Mixup learning strategies for text-independent speaker verification. In *Interspeech*, pages 4345–4349, 2019.