



Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model

Tarun Gupta¹, Duc-Tuan Truong², Tran The Anh², Chng Eng Siong²

¹Department of Computer Science and Engineering, Indian Institute of Technology, Indore, India

²School of Computer Science and Engineering, Nanyang Technological University, Singapore

cse1800001059@iiti.ac.in, {DUCTUAN001, TRANTHEA001}@e.ntu.edu.sg, ASESChng@ntu.edu.sg

Abstract

The estimation of speaker characteristics such as age and height is a challenging task, having numerous applications in voice forensic analysis. In this work, we propose a bi-encoder transformer mixture model for speaker age and height estimation. Considering the wide differences in male and female voice characteristics such as differences in formant and fundamental frequencies, we propose the use of two separate transformer encoders for the extraction of specific voice features in the male and female gender, using wav2vec 2.0 as a common-level feature extractor. This architecture reduces the interference effects during backpropagation and improves the generalizability of the model. We perform our experiments on the TIMIT dataset and significantly outperform the current state-of-the-art results on age estimation. Specifically, we achieve root mean squared error (RMSE) of 5.54 years and 6.49 years for male and female age estimation, respectively. Further experiment to evaluate the relative importance of different phonetic types for our task demonstrate that vowel sounds are the most distinguishing for age estimation.¹

Index Terms: speaker profiling, age estimation, height estimation, mixture of experts.

1. Introduction

Speech is an acoustic output produced by precisely coordinated movement of different human body parts. Therefore, there have been suggestions that acoustic features of speech can convey information about the physical characteristics of the speaker. Among various physical characteristics, scientific studies have investigated the correlation between voice characteristics and a speaker age and height. Authors of [1, 2] reported that the vocal tract length, sub-glottal resonance frequencies, and formant frequencies are correlated with the individual's height. Other voice characteristics of speech such as speech rate, sound pressure level, fundamental frequency, etc. vary according to the speaker's age [2, 3]. The age-related glottis deterioration of the speaker also impacts the speech characteristics like jitter, shimmer [4], and speech harmonics [5].

Automatic speaker profiling systems could be applied to a variety of different fields. For example, in criminal investigation, evidence could be in the form of voice recordings, e.g. a hoax bomb threat or ransom demand over a phone call [6, 7]. In such cases, estimating the age and height of speakers in the audio evidence could save investigating agencies' time by narrowing down the number of suspects. Similarly, predicting the age and gender of a speaker from the speech data can aid marketing campaigns by targeting suitable gender/age customer

groups [7]. In addition, the speaker profiling system can benefit other speech fields, namely, speaker diarization and speech-based verification as well.

Contribution: In this work, we study the use of self-supervised speech representation, specifically wav2vec 2.0 [8], for speaker age and height estimation. We also compare wav2vec 2.0 to other famous feature representations: MFCC, filter bank, and x-vectors [9]. To our best knowledge, our work is the first in the literature to demonstrate the potential of wav2vec 2.0 in predicting speaker age and height. For the downstream architecture, we present a novel Mixture of Experts (MoE) based bi-encoder transformer model that utilizes wav2vec 2.0 as a common-level speech feature extractor followed by a bi-encoder architecture. The bi-encoder architecture is motivated from the differences in male and female voice characteristics such as differences in formant and fundamental frequencies [10, 11]. The two encoders act as 'experts' for providing distinctive features for male and female voice. We further make use of homoscedastic uncertainty [12] to define the multi-task loss function and mixup [13] as a regularization technique. The proposed network achieve new state-of-the-art results in age estimation on the TIMIT dataset.

The rest of our paper is organized as follows. Section 2 briefly describes the speaker profiling literature. Section 3 illustrates our proposed model architecture and techniques. Section 4 describes the experiments conducted to estimate the age and height of a speaker in a monolingual setting using the TIMIT dataset. Also, this section discusses the impact of the different phones on the estimation result in Section 4.3. Finally, conclusions are presented in Section 5.

2. Related works

There is a wide choice of speech feature extractors for automatic height and age estimation available in the literature. However, most of the previous studies used conventional techniques of extracting features from raw speech signals. In the earlier years, authors of [15, 16] used the Open-Smile toolkit to convert the short-term spectral features into various statistics like mean, median, percentiles, etc. for height and age estimation. A similar statistical approach for age and height prediction applied i-vector (dimension reduced version of Gaussian mixture model universal background model) [17, 18] to convert a variable-length utterance into a fixed-size embedding vector. Another embedding approach of speaker profiling is obtaining spectral characterizations of the speech signal. For example, the speaker's resonance frequencies of sub-glottal are used for height estimation [19]. Singh et al. [2] applied a bag of words representation to capture short-term cepstral features with different time resolutions. Other common-level short term features are Mel Frequency Cepstral Coefficients (MFCC) [20, 21], cep-

¹Code and models are available at github.com/tarun360/SpeakerProfiling

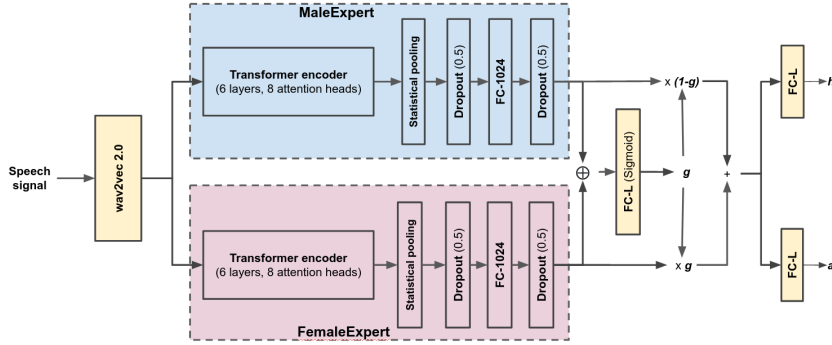


Figure 1: *wav2vec 2.0* bi-encoder model architecture. Transformer encoder [14] with 6 layers and 8 attention heads is used. Statistical pooling refers to taking mean and standard deviation along frame dimension and concatenating them to obtain utterance level representation. ‘FC-L’ denotes a fully connected layer with L neurons. \oplus denotes the concatenation operation. a , h and g denote the age, height and gender prediction.

stral and pitch features [4, 22], etc.

In this era of deep learning, deep neural networks (DNN) have shown the outstanding ability to discover descriptive and distinctive representations from raw speech audio. Therefore, recent studies have applied DNN-based speech representation learning to improve the accuracy of the speaker’s height and age prediction. Abumallouh et al. [23] presented a speaker age and gender classifier that based on top of an unsupervised DNN bottleneck feature extractor can achieve better performance than its original MFCCs feature set, especially for female speaker. By applying DNN discriminative embedding called x-vector [24], authors of [25, 26] gained better results in height and age estimation in TIMIT dataset compared to previous studies. Shangeth et al. [27] employed semi-supervised learning approach to learn speaker representation and achieve the state-of-the-art result on age estimation on the TIMIT test set with Mean Absolute Error (MAE) of 4.8 and 5.0 years for male and female speakers, respectively. Recently, Baevski et al. [8] presented the *wav2vec 2.0* framework for self-supervised learning of discrete speech units. In the field of speaker recognition, *wav2vec 2.0* have increased the classification accuracy considerably since it can capture much more phonetics information than its conventional counterpart. To the best of our knowledge, there is no work in the literature which studies the use of *wav2vec 2.0* for speaker height and age estimation.

Literature also showed that there are differences between male and female voices. The average female formant and fundamental frequencies are higher than that of the male speaking voice [10, 11]. Therefore, the extraction of height and age information in the speech signal is also affected by the gender of the speaker [28]. Most of the previous speaker profiling studies considered gender classification and height/age estimation as a joint problem [26, 27]. The only work to use gender information and get a slight improvement in speaker’s height and age prediction is by [22, 29]. Nevertheless, these works adapted ground truth gender information which is fed to their model just as a binary value. In order to capture distinctive representations of male and female speech, we use two experts, one for each gender. Mixture of Experts (MoE) [30] is a supervised learning system in which separate networks are designed to handle different subsets of the training samples. Recently, the MoE architecture has been studied for various tasks in the speech domain such as multi-accent speech recognition [31], code-switching speech recognition [32], etc.

3. Method

3.1. Self supervised representation

Motivated by the success of self-supervised learning in the field of speech recognition, we explore an SSL model, namely *wav2vec 2.0* [8] in speaker profiling. The *wav2vec 2.0* model learns speech representations by solving contrastive tasks in latent space. Specifically, *wav2vec 2.0* is composed of 1D convolution features extractor $f : X \rightarrow Z$ which expects raw audio input X and outputs latent representations Z , followed by transformer encoders $e : Z \rightarrow C$, which provides contextual information. The feature extractor outputs are quantized by a quantization module.

3.2. Mixture of Experts

The Mixture of Experts (MoE) [30] concept is based on divide-and-conquer principle. If the training dataset is known in advance to be divided naturally into certain subspaces, then separated experts of these subspaces can be trained. A gating network decides the weights to be assigned to each of the expert views and then a weighted sum of all the expert views is taken. The MoE architecture leads to less interference during back-propagation which may lead to better generalizability.

3.3. Model architecture

The model architecture has been described in Fig.1. First, *wav2vec 2.0* is used to extract features from the raw audio waveform. Then, utilizing the MoE concept, a bi-encoder transformer network is built on top of the extracted features. For the two genders present in the TIMIT dataset, male and female, we consider separate experts, *MaleExpert* and *FemaleExpert*. These two experts have identical architectures as illustrated in Fig. 1, having 6 layers of self-attention based transformer encoder with 8 attention heads in each layer [14]. We take an average and standard deviation of the output of the transformer encoder along the frame dimension and concatenate them to get utterance level representations. These utterance level representations are then fed to dropout and fully connected layers to complete the expert architecture. The two separate experts can focus on gender-based audio characteristics important for estimating height and age, while the fine-tuned *wav2vec 2.0* can learn common-level features for both genders.

The two experts, *MaleExpert* and *FemaleExpert* pro-

Table 1: Comparison of the proposed wav2vec 2.0 bi-encoder model with the existing work.

Method	Height RMSE		Height MAE		Age RMSE		Age MAE	
	Male	Female	Male	Female	Male	Female	Male	Female
Singh et al. [2]	6.7	6.1	5.0	5.0	7.8	8.9	5.5	6.5
Kalluri et al. [21]	6.85	6.29	-	-	7.60	8.63	-	-
Kwasny et al. [25]	-	-	-	-	7.24	8.12	5.12	5.29
Williams et al. [33]	-	-	5.37	5.49	-	-	-	-
Mporas et al. [15]	6.8	6.3	5.3	5.1	-	-	-	-
Shangeth et al. (single-task model) [27]	8.1	6.0	5.9	4.9	6.96	7.6	4.8	5.1
Shangeth et al. (multi-task model) [27]	7.5	6.5	5.8	5.1	6.8	7.4	4.8	5.0
Manav et al. (single-task model) [29]	6.92	6.24	5.20	4.95	7.20	7.10	5.04	5.02
Manav et al. (multi-task model) [29]	6.95	6.44	5.26	5.15	7.81	8.60	5.50	5.89
wav2vec 2.0 bi-encoder (ours)	7.3	6.43	5.58	5.07	5.54	6.49	3.96	4.48

Table 2: Comparison of wav2vec 2.0 bi-encoder model with wav2vec 2.0 single-encoder model.

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

vide two expert views, e_m and e_f :

$$e_m = \text{MaleExpert}(x) \quad (1)$$

$$e_f = \text{FemaleExpert}(x) \quad (2)$$

where x is the feature representation extracted from the fine-tuned wav2vec 2.0. e_f and e_m are concatenated and fed to a fully connected layer with sigmoid activation to predict the gender $g \in [0, 1]$. Male gender is encoded as 0 and female as 1. Gender prediction acts as a gating network to combine the output of the two experts as follows:

$$e = (1 - g) \times e_m + g \times e_f \quad (3)$$

e is then fed to fully connected layers to perform age and height regression.

3.4. Loss function

In our multi-task model, three losses are factored into the training process, namely height, age and gender loss. Prior approaches for building a multi-task model for speaker profiling have used a weighted sum of these losses [27, 29], where the loss weights are manually fine-tuned. Instead, we use uncertainty loss [12], which uses homoscedastic uncertainty to combine multiple losses. Using this, we define our multi-task loss function as:

$$\mathcal{L} = \frac{\mathcal{L}_{\text{height}}}{2\sigma_{\text{height}}^2} + \frac{\mathcal{L}_{\text{age}}}{2\sigma_{\text{age}}^2} + \frac{\mathcal{L}_{\text{gender}}}{2\sigma_{\text{gender}}^2} + \log(\sigma_{\text{height}}\sigma_{\text{age}}\sigma_{\text{gender}}) \quad (4)$$

where σ_{height} , σ_{age} and σ_{gender} are learnable parameters. As suggested in their paper, we make the substitution $s_{\text{height}} = \log(\sigma_{\text{height}}^2)$, $s_{\text{age}} = \log(\sigma_{\text{age}}^2)$ and $s_{\text{gender}} = \log(\sigma_{\text{gender}}^2)$ in equation 4 for numerical stability.

3.5. Mixup

We use mixup as a regularization technique to encourage the model to learn linear inference from the input audio. Mixup

augmentation has been used previously in text-independent speaker verification [34].

Given two audio samples x_i, x_j , with h_i, h_j, a_i, a_j and g_i, g_j as the corresponding height, age, and gender values, the mixup augmented sample is defined as:

$$x_{\text{mixup}} = \lambda x_i + (1 - \lambda)x_j \quad (5)$$

$$h_{\text{mixup}} = \lambda h_i + (1 - \lambda)h_j \quad (6)$$

$$a_{\text{mixup}} = \lambda a_i + (1 - \lambda)a_j \quad (7)$$

$$g_{\text{mixup}} = \lambda g_i + (1 - \lambda)g_j \quad (8)$$

where $\lambda \sim U(0, 1)$. To deal with audio of different lengths, we repeat the audio of a shorter length to make it of the same length as the longer audio.

4. Experiments

4.1. Dataset

In this work, we use TIMIT dataset [35] for our experiments. It consists of audio samples from 630 speakers, having eight different dialects of American English. The height varies in the range of 145cm to 204cm, and age varies in the range of 21 years to 76 years. TIMIT dataset comes pre-divided into the train and test set and we randomly select 15% of the train set for validation.

4.2. Experimental Setup

For the proposed bi-encoder transformer model, apart from wav2vec 2.0, we also consider other features for comparison: MFCC, filter bank and x-vectors. For filter bank, we consider 80 mel bins along with first and second order delta features. For MFCC, we consider 16 cepstral coefficients, along with first and second order delta features. For both filter bank and MFCC, we use apply Cepstral Mean and Variance Normalization (CMVN)

Table 3: Comparison of different feature extractors: MFCC, filter bank, x-vectors and wav2vec 2.0.

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

Table 4: Percentage change in RMSE values due to phoneme masking

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

and use frame length of 25ms and frame shift of 10ms. For x-vectors, we extract the frame-level features before its statistical pooling layer, which is then fed to downstream bi-encoder architecture. The x-vectors were pre-trained on VoxCeleb1 [36] and VoxCeleb2 [37] training data.

In another experiment, to show the efficacy of bi-encoder architecture, we compare the wav2vec 2.0 bi-encoder model with wav2vec 2.0 single-encoder model. The wav2vec 2.0 single-encoder model has same layers as the wav2vec 2.0 bi-encoder model, except for the fact that there is only a single expert for both genders and there is no gating network. The output of the single expert is fed to fully connected layers to perform age, height, and gender estimation.

For models using wav2vec 2.0, we unfreeze the entire wav2vec 2.0 architecture except for the first five convolution layers of the convolutional feature extractor and use Adam optimizer [38] with the learning rate of 10^{-6} . When using MFCC, filter bank and x-vectors, we use Adam optimizer with the learning rate of 10^{-5} . Mixup was proposed as regularization technique for deep neural network architectures to favor simple linear behavior, and as such, cannot be applied directly at audio-level for traditional feature extraction techniques like MFCC and filter bank. As a result, we use mixup only when utilizing x-vectors and wav2vec 2.0 as common feature extractor.

4.3. Experimental Results

In Table 1, we compare the results of the wav2vec 2.0 bi-encoder model with previous works. While many previous works built separate models for age and height, or separate models for male and female, we compare our multi-task model with all of them. We report our results both in root mean squared error (RMSE) and mean absolute error (MAE). It can be observed that the model achieves significant improvement in the age estimation task. Specifically, we achieve RMSE error 5.54 years and 6.49 years, which corresponds to relative improvement of 18.5% in male age estimation and 8.6% in female age estimation over the current state-of-the-art.

The comparison of wav2vec 2.0 bi-encoder vs wav2vec 2.0 single encoder model has been illustrated in Table 2. The wav2vec 2.0 bi-encoder model achieves a relative improvement of 2.9% for male age estimation and 7.0% for female age estimation over wav2vec 2.0 single-encoder model in terms of RMSE error, testifying our hypothesis of using separate encoders for the two genders. However, the same pattern is not observed in height estimation results, indicating that the bi-encoder architecture is not as useful for height estimation as it is for age estimation.

In Table 3, we tabulate the results of different feature extractors: MFCC, filter bank, x-vectors and wav2vec 2.0. It can be observed that wav2vec 2.0 feature representation is superior to other feature extraction techniques.

4.4. Phonetic importance analysis

In order to understand the relative importance of the different types of phones in speaker profiling, we analyze the results of wav2vec 2.0 bi-encoder model after masking all utterances of a particular phone type. TIMIT dataset provides time-aligned phonetic transcriptions for each audio sample. Phones used in TIMIT corpus are divided into the following classes: ‘Stops’, ‘Affricates’, ‘Fricatives’, ‘Nasals’, ‘Semivowels and Glides’, ‘Vowels’ and ‘Others’. For each of these phone types, we mask all the phones in audio samples in TIMIT test set, and then calculate the height and age RMSE scores. The relative percentage change due to these phone masking, as compared to when no masking was done, are tabulated in Table 4. We notice the largest increase in age RMSE value due to ‘Vowel’ masking, implying vowel sounds are the most important phoneme type for age estimation. Surprisingly, we notice no significant change in height RMSE values, indicating that height estimation is not dependent on any particular phoneme type.

5. Conclusions

In this work, we present a Mixture of Experts (MoE) based bi-encoder transformer model, that uses self-supervised representation, specifically wav2vec 2.0, for speaker age and height estimation. The experimental results demonstrate that having separate experts for male and female voices can reduce interference during training process and can achieve state-of-the-art results for age estimation. We employed the homoscedastic uncertainty principle to combine multiple losses of our multi-task model. As part of future work, we plan to explore different feature extractors and self-supervised learning to further improve age and height estimation results.

6. Acknowledgements

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (<https://www.nsc.sg>).

7. References

- [1] H. Arsikere, S. M. Lulich, and A. Alwan, "Estimating speaker height and subglottal resonances using mfccs and gmms," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 159–162, 2014.
- [2] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *2016 4th International Conference on Biometrics and Forensics (IWBF)*. IEEE, 2016, pp. 1–6.
- [3] S. Schötz, "Acoustic analysis of adult speaker age," in *Speaker classification I*. Springer, 2007, pp. 88–107.
- [4] C. Müller and F. 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.
- [5] M. Li, K. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer, Speech, and Language*, vol. 27, 11 2012.
- [6] R. Singh, J. Keshet, and E. Hovy, "Profiling hoax callers," in *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, 2016, pp. 1–6.
- [7] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, pp. 4–39, 2013.
- [8] A. Baeovski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [10] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.
- [11] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (ltas)," *Journal of Voice*, vol. 10, no. 1, pp. 59–66, 1996.
- [12] A. Kendall, Y. Gal, and R. 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*, 2018, pp. 7482–7491.
- [13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [15] I. Mporas and T. Ganchev, "Estimation of unknown speaker's height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.
- [16] T. Ganchev, I. Mporas, and N. Fakotakis, "Audio features selection for automatic height estimation from speech," in *Hellenic Conference on Artificial Intelligence*. Springer, 2010, pp. 81–90.
- [17] S. O. Sadjadi, S. Ganapathy, and J. 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)*, 2016, pp. 5040–5044.
- [18] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on lstm recurrent neural networks," *IEEE Access*, vol. 6, pp. 22 524–22 530, 2018.
- [19] H. Arsikere, G. K. F. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation," *Speech Commun.*, vol. 55, pp. 51–70, 2013.
- [20] B. Pellom and J. 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*, vol. 2, 1997, pp. 873–876 vol.2.
- [21] S. B. Kalluri, D. Vijayasenan, and S. 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)*. IEEE, 2019, pp. 6580–6584.
- [22] M. Kaushik, V. T. Pham, and E. S. Chng, "End-to-end speaker height and age estimation using attention mechanism with lstm-rnn," *arXiv preprint arXiv:2101.05056*, 2021.
- [23] A. Abumallouh, Z. Qawaqneh, and B. Barkana, "New transformed features generated by deep bottleneck extractor and a gmm-ubm classifier for speaker age and gender classification," *Neural Computing and Applications*, vol. 30, 10 2018.
- [24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [25] D. Kwasny and D. Hemmerling, "Joint gender and age estimation based on speech signals using x-vectors and transfer learning," *arXiv preprint arXiv:2012.01551*, 2020.
- [26] D. Kwaśny and D. Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors (Basel, Switzerland)*, vol. 21, 2021.
- [27] S. Rajaa, P. Van Tung, and C. E. Siong, "Learning speaker representation with semi-supervised learning approach for speaker profiling," *arXiv preprint arXiv:2110.13653*, 2021.
- [28] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Communication*, vol. 121, pp. 16–28, 2020.
- [29] M. Kaushik, T. T. Anh, E. S. 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)*. IEEE, 2021, pp. 1–8.
- [30] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [31] A. Jain, V. P. Singh, and S. P. Rath, "A multi-accent acoustic model using mixture of experts for speech recognition," in *INTERSPEECH*, 2019, pp. 779–783.
- [32] Y. Lu, M. Huang, H. Li, J. Guo, and Y. Qian, "Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts," in *INTERSPEECH*, 2020, pp. 4766–4770.
- [33] K. A. Williams and J. H. Hansen, "Speaker height estimation combining gmm and linear regression subsystems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7552–7556.
- [34] Y. Zhu, T. Ko, and B. Mak, "Mixup learning strategies for text-independent speaker verification," in *Interspeech*, 2019, pp. 4345–4349.
- [35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.
- [36] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Interspeech 2017*, Aug 2017.
- [37] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, Sep 2018.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.