End-to-End Speaker Age and Height Estimation using Attention Mechanism and Triplet Loss

Manav Kaushik*, Van Tung Pham†, Tran The Anh†, Eng Siong Chng†
* Birla Institute of Technology and Science (BITS) Pilani
E-mail: f2016472@pilani.bits-pilani.ac.in
† Nanyang Technological University (NTU), Singapore
E-mail: vtpham@ntu.edu.sg, theanh.tran@ntu.edu.sg, ASESChng@ntu.edu.sg

***Abstract*—Automatic age and height estimation of speakers using acoustic features is widely used for the purpose of human-computer interaction, forensics, etc. In this work, we study end-to-end framework for age and height estimation. We first propose a novel attention mechanism, named cross-attention. Different from conventionally used attention, which calculates context vector as the sum of attention only across timeframes, the proposed approach introduces a modified context vector which takes into account total attention across both time-frames and encoder units. We further propose using triplet loss to enhance the discriminative power of the encoder.**

We evaluate the Root Mean Square Error (RMSE) of proposed approaches on the TIMIT corpus. The proposed cross-attention outperforms the conventional counterpart for both age and height estimation while the triplet loss brings 8% relative improvement for age estimation. We obtain RMSE of 6.92/6.42 cm for male/female height estimation and 7.20/7.10 years for male/female age estimation which outperforms all the previous baselines achieving state-of-the-art performance for age estimation on TIMIT dataset. By tracking the attention weights allocated to different phones, we find that Vowel phones are most important while Stop and Fricative phones are least important for the estimation task.

Index Terms: Automatic height and age estimation, multi-task learning, Attention, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Triplet Loss.

I. Introduction

Speech is a unique physiological signal which not only contains information about the linguistic content (such as words, accent, language, etc.) but also conveys the para-linguistic content (such as height, age, gender, emotions, etc.). This helps us in estimating the physical parameters like height and age of a speaker, which holds a wide variety of applications in the real-world such as natural human-machine interaction, speaker profiling, and forensics [1], [2]. Although speech signals help us to estimate several speaker characteristics, we limit ourselves to only age and height estimation of speakers using speech signals for the purpose of this study.

A typical approach for speaker characteristic estimation is to apply shallow learning techniques, such as linear regression [3] or support vector machine [4], [5], [6], on top of utterance-level representation such as i-vector [4], [6] or x-vector [7]. Such approaches are not end-to-end since the utterance-level representation extractors are trained separately for speaker recognition tasks which are not optimized for height and age estimation.

Recently, end-to-end approach has been studied for age and height estimation and produces better results than the traditional approach [8]. In this work, we not only follow this approach but also improve it by introducing several novelties.

Firstly, we propose a novel soft-attention mechanism for speaker characteristic estimation task. As from our best knowledge, there is not any work in the literature which studies the use of attention for speaker age and height estimation. More importantly, instead of performing attention across speech frames, as done conventionally [9], [10], we perform attention across both speech frames and encoder units to obtain two context vectors and then combine them to generate a final context vector. We believe that the proposed attention, denoted as cross attention, captures more information than the conventional counterpart and hence could produce better performance. Thus, our motivation to modify the conventionally used attention mechanism is to exploit the information that may be captured across another dimension, i.e. across the encoder units, in order to better estimate the characteristics of a speaker.

Secondly, to enhance the discriminative power of the encoder, we propose using Triplet Loss [12], [13], [14] in combination with Mean Squared Error Loss during training the speaker profiling systems. Triplet loss enforces the encoder to produce embeddings that have larger inter-class variation and smaller intra-class variation, which results in better estimation. Note that previous works used Triplet Loss with classification problems while our task is regression task, thus, we need to perform quantization on our data. We convert continuous age and height labels into discrete classes so as to train the embeddings to cluster around their own class and away from other classes using triplet loss. Our primary motivation for studying triplet loss is the fact that length of vocal tracts and glottal-pulse rate (variance rate in voice quality affected by the changes in the folds of the vocal cords during sound utterance) vary with age and height [15], [16], [17]. For example, taller people may be expected to have longer vocal tracts. Hence, the vocal tract resonances present in the speech signal may be expected to provide us information about a speaker's height or age. In order to capture the variation of vocal tract's length and structure with varied age and height, we try to obtain embeddings from cross attention layer, after training with triplet loss, so as to make the model capable of differentiating and clustering speech samples on the basis of age or height classes.

Lastly, by analyzing attention weights across speech frames, we find that that highest weights have been assigned to Vowel phones while the lowest weights have been assigned to Stop phones and some of the Fricative phones. Since speaker characteristics such as height and age are correlated with the length of speaker vocal tract and glottal-pulse rate [17], this higher attention to vowel phones maybe attributed to the fact that these phones involves significant vibrations of the vocal tract folds, while utterance of stop phones do not involve such vocal vibrations, instead, their production requires a complete closure of the vocal tract which may be a reason for their lower attention weights. Lesser attention for Fricative phones may be justified because during their utterance, the vocal folds do not vibrate and the glottis remains open with continuous airflow which does not carry much information.

Apart from above contributions, we also study how passing in gender information as a feature helps the model to better estimate the height and age of a speaker.

We have organized the paper in the following format: Section II discusses the related works from the litearture followed by Section III which describes the dataset used for our experiments. Section IV explains all the techniques that we adopted for our work and Section V describes the experimental setup that we followed and the consequent results that we obtained with proper comparison with other works. Finally, in Section VI, we conclude our work.

II. Related Works

Most of previous studies on height and age estimation tend to use conventional approaches of applying shallow learning techniques. For instance, Williams et al. [3] combine Gaussian Mixture Models (GMM) and linear regression subsystems to estimate the speaker height. Poorjam et al. [4] and Bahari et al. [18] predict speaker height and age by applying least-squares Support Vector Regression (SVR) on top of i-vector. Mahmoodi et al. [5] use Support Vector Machines (SVMs) while Bocklet et al. [6] use GMM supervectors with SVM for age estimation task. Singh et al. [16] use a bag of words representation generated from short-term cepstral features and train a Random Forest regressor for age and height estimation. The issue with the above mentioned approaches is that none of them are end-to-end modeling techniques and thus, are not specifically optimized to speaker physical parameter estimation such has height and age.

More recently, Ghahremani et al. [7] propose an end-to-end deep neural network (DNN) for age prediction while Kalluri et al.[8] also attempt to jointly predict both height and age of speaker using a unified end-to-end DNN model which is initialized using a conventional system based on SVR trained with Gaussian Mixture Model-Universal Background Model (GMM-UBM) supervector features. Although, both of these works employ an end-to-end architecture of their estimation tasks, they rely more on conventional approaches.

The Attention mechanism was proposed by Bahdanau et al. [19] who use this mechanism for the task of neural machine translation. Shan et al. [10] use soft attention mechanism for key-word spotting task. Apart from this, Attention has found successful application in computer vision tasks as well such as object recognition and image captioning [20], [21], [9]. Such success of attention mechanism encouraged us to explore its utility for our task as well.

Initially motivated by Weinberger et al. [12] in the context of nearest-neighborhood classification, Triplet Loss was successfully used by Schroff et al. [13] to train a convolutional neural network (CNN) to learn an embedding for faces. Apart from this, Ding et al. [22] also employ a triplet loss to get the relative distance between images for person re-identification.

To our best knowledge, none of the past works in the literature have used attention or triplet loss for speaker physical parameter estimation. Our work is the first in the literature to demonstrate the potential of attention mechanism and triplet loss in tracking the relational dependency and importance of different phones in estimating speaker height and age in an utterance.

III. Dataset Used

We use the TIMIT dataset [23] for all our experiments done in this study. TIMIT has a total of 6300 unique utterances. There are 630 speakers of these utterances who are distributed across 8 different dialect regions with each speaker speaking ten different utterances. The gender distribution of the speakers in male to female is 2:1. Moreover, the dataset also includes time-aligned orthographic, phonetic and word transcriptions which help us track phonetic attention.

The train-test split is given in the dataset i.e. 461 speakers (326 male and 135 female) for training and validation, and 162 speakers (112 male and 56 female) for testing. The height of speakers in the training data ranges from 145cm to 199cm and in testing data, they range from 153cm to 204cm. Similarly, the age of speakers ranges from 21 years to 76 years in training data and 22 years to 68 years in test data. There is no overlapping of speakers between test and training datasets. Moreover, the duration of the utterances ranges from 1- 6s with an average of about 2.5s.

IV. Methodologies

Our proposed framework for end-to-end speaker height and age estimation is shown in Fig. 1 with the output shape of each layer mentioned beneath the layer's name. First, we obtain Filter bank energies and pitch features from raw data and preprocess them before passing as inputs to our models. Then these input features are encoded by an LSTM network before feeding them to an attention layer. Subsequently, the output of attention layer, which is a vector, is transformed by a dense layer to make age and height predictions.

For our final model, we use triplet loss (as shown in Fig. 2) on the embeddings obtained from the cross attention layers after passing these embeddings through an L2 Normalization process. It maybe noted that triplet loss is primarily used for classification tasks while our predictions are regressive, thus, we first quantize the age and height labels. More specifically, we convert continuous age and height labels into discrete classes so as to train the embeddings to cluster around their own class and away from other classes using triplet loss.

In following subsections, we describe these processes in detail.

A. Data Preprocessing and Augmentation

Our input for each utterance is a two-dimensional matrix consisting of T time-frames with each timeframe consisting of 83 acoustic features (80 filter bank and 3 pitch features) extracted from windows of 25ms with 10ms stride. We apply Cepstral Mean and Variance Normalization (CMVN) to these acoustic features. The resulting features are $\mathbf{x} = [x_1, x_2, \ldots, x_T]$ where each frame, x_i, contains 83 features.

Since neural network models require a huge amount of data to train properly, we use speed perturbation as an augmentation step to obtain the audio signals at 1.1x and 0.9x speeds as well. Apart from this, we use spectral augmentation (SpecAugment) [24] to enhance the robustness of our model by randomly masking strands from feature and time axes covering approximately 15-20% of the training data.

B. LSTM RNN Encoding

Although deep neural networks (DNNs) have successfully been used for numerous speech related tasks, they primarily rely on stacking a few frames to account for the context of the speech sample and this restricts any long-term dependency and results in models with a much bigger number of free parameters that need to be trained for the same data [25]. Recurrent neural networks (RNNs) are deep neural networks (DNNs) that solve this problem as they have connections which form a directed cycle and thus, can exhibit dynamic temporal behavior. However, classical RNNs suffer from the problem of vanishing gradient [26] which implies that the gradient of the output error with respect to previous inputs quickly vanishes as the time lags between relevant inputs. Long Short-Term Memory (LSTM) recurrent neural network replaces the hidden units in a RNN with memory blocks allowing the LSTM to capture long term dependency storing temporal state of the network so the output depends not only on the input but also on previous inputs.

Since LSTM has shown to be efficient to capture long temporal dependencies, we choose this architecture to encode acoustic features for our study. Given a sequence of input features $\mathbf{x} = [x_1, x_2, \ldots, x_T]$, the LSTM network processes it frame-by-frame to generate the sequence of hidden states $\mathbf{h} = [h_1, h_2, \ldots., h_T]$ where each state h_i has dimension of n_{units} i.e. number of LSTM units.

Once the input features are encoded, the straightforward approach is to take the final hidden state i.e. h_T as the utterance-level representation for height and age estimation. However, in practice, LSTM tends to forget information when operated on longer sequences. Therefore, we propose to use attention mechanism to solve this problem as presented in Section IV-C.

C. Attention Mechanism

The attention mechanism was primarily introduced to help memorizing long sentences in neural machine translation [19]. It generates a context vector as weighted sum of hidden states of the LSTM encoder over all timeframes. Since the attention mechanism has access to the entire input sequence, the problem of forgetting initial parts of the sequence is solved. We use soft attention which is typically used in previous works [9], [10]. It maybe noted that we do not employ the popular self attention or multi-head attention as they are intra-sequence attention (i.e. attention relating different positions of a single sequence in order to compute a representation of the same sequence) while our task is predictive in nature. Multi-head attention is primarily used in tasks such as text summarization, image description and machine reading [11].

First, a scalar score e_t is estimated for each LSTM hidden state h_t as:

$$e_t = \mathbf{v_a}^\intercal tanh(\mathbf{W_a} h_t + \mathbf{b}) \tag{1}$$

where $\mathbf{v_a}$, $\mathbf{W_a}$ and $\mathbf{b}$ are learnable parameters. Then, the attention weights α_t are obtained by applying a softmax function on e_t, i.e.

$$\alpha_t = \frac{exp(e_t)}{\sum_{i=1}^{T} exp(e_i)} \tag{2}$$

Since we use a softmax function, $\alpha_t \epsilon [0, 1]$ and $\sum_{t=1}^{T} \alpha_t = 1$. After this, we obtain a context vector, $\mathbf{c}$, as the weighted average across all timeframes of the LSTM outputs $\mathbf{h}$:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t h_t \tag{3}$$

As a result, $\mathbf{c}$ has the same dimension as hidden states h_t i.e. n_{units}

Instead of considering only $\mathbf{c}$ as the final context vector (which is the conventional approach [9], [10]), we propose a cross-attention approach in which we further perform attention across all the n_{units} LSTM units to generate another context vector, denoted as $\mathbf{c}^*$, and concatenate them to obtain the final context vector $\mathbf{f}$.

$$\mathbf{f} = [\mathbf{c}, \mathbf{c}^*] \tag{4}$$

Note that $\mathbf{c}^*$ has dimension of n_{frames}, hence $\mathbf{f}$ has dimension of (n_{frames} + n_{units}). The mechanism performing attention across both speech frames and encoder units to capture more information is expected to produce better performance.

$\mathbf{f}$ is finally passed into a dense layer which makes the final prediction.

D. Dense Layers

For each of height or age, the estimation made as:

$$\hat{y} = ReLU(\mathbf{v}^\intercal \mathbf{f}) \tag{5}$$

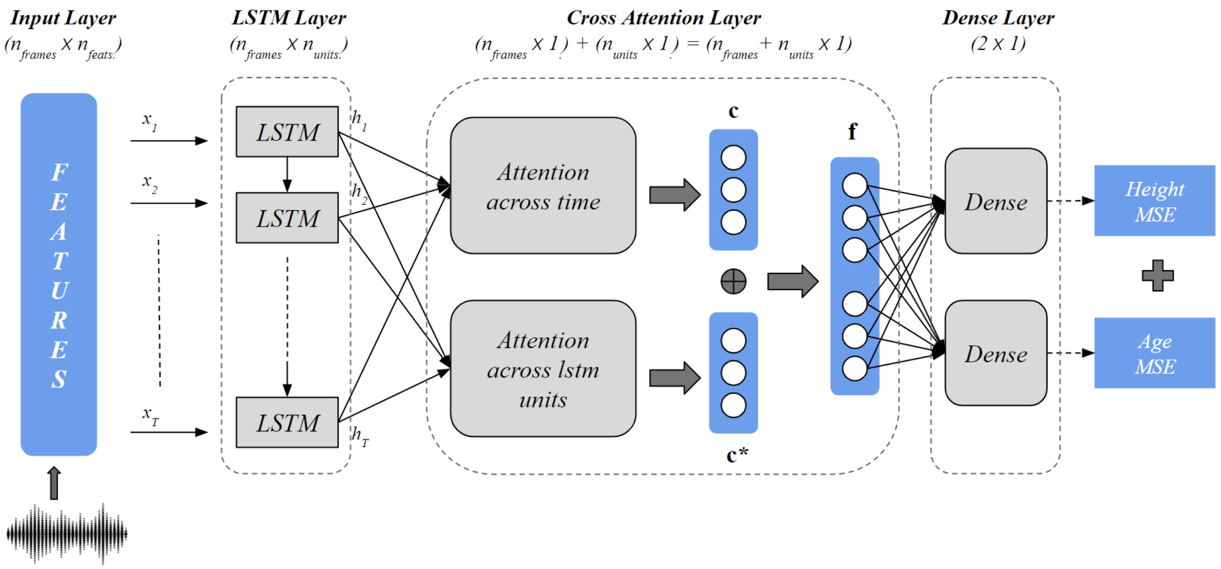

Fig. 1. Proposed Cross-Attention + Multitask Learning Model for simultaneous age and height estimation

where **v** is a learnable vector of length same as **f**. We use Mean Squared Error (MSE) as our loss function:

$$loss = \frac{1}{N}[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2] \qquad (6)$$

where y_i and $\hat{y}_i$ are the actual and predicted values respectively for utterance i and N is the total number of utterances.

We also study multi-task learning which aims to estimate height and age at the same time. The training loss for our approach is calculated as:

$$Loss_{total} = a(loss_{height}) + (1-a)(loss_{age}) \qquad (7)$$

where a is a hyper-parameter optimized on the validation set.

E. Triplet Loss

The intuition behind employing triplet loss [13] is to learn a Euclidean embedding per utterance using our LSTM-Cross-Attention framework as age and height variations have shown to be significantly correlated to vocal tract length and glottal-pulse rate which may be captured using discriminatively training embeddings, obtained from the cross attention layer, using triplet loss. The network is trained such that the squared L2 distances in the embedding space directly correspond to height/age class similarity i.e. utterances from the same class have smaller distance while those from distinct classes have larger distances. Here the classes are formed based on the height/ age group of the speaker. For instance, as shown in Figure 2(a), utterances having speaker with height 140cm to 145cm belong to $class\_0$, 145cm to 150cm belong to $class\_1$ and so on. Similarly, for age, utterances having speaker with age in the range of 20 years to 25 years belong to $class\_0$, 25 years to 30 years belong to $class\_1$ and so on.

The embedding is represented by $f(x)$. Here, we want to ensure that an utterance x_i^a (anchor) of a specific class is closer to all other utterances x_i^p (positive) of the same person than it is to any utterance x_i^n (negative) belonging to a different class [13]. Thus, the loss function to be minimized is as follows:

$$TL = \sum_1^N [||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2] \qquad (8)$$

where utterance x_i^a is anchor, x_i^p is positive of the same class, x_i^n is negative belonging to a different class.

V. Experimental Results

A. Experimental Setup

The raw inputs for our model are speech samples, all of which are sliced or padded to attain 8 seconds in length. These raw inputs are used to obtain filter bank and pitch features which are then preprocessed as discussed in IV-A. As a result, each input sample for our model is of the shape: (800 × 83) where, 800 represents the number of time-frames i.e. n_{frames} and 83 represents the number of features per frame i.e. n_{feats}. When gender is passed in as a binary feature (i.e. 0 or 1) in the input layer, its shape becomes (800 × 84). The LSTM network consists of a single layer LSTM with 64 units. Thus, the output of the LSTM layer is (800×64) where 64 represents the number of lstm units i.e. n_{units}. We use a recurrent dropout of 20% to avoid overfitting in the LSTM layer. This encoded output of the LSTM is passed into the cross attention layer which creates two context vectors: **c** (across time-frames) with shape (64×1) and **c*** (across lstm units) with shape (800×1). These two context vectors are concatenated to give us the final context vector, **f**, of shape (864 × 1) where 864 represents dimension of the cross attention context vector i.e. (n_{frames} + n_{units}). And finally, this context vector is passed into the Dense with single output for single-task learning (height or age) and two outputs for multi-task learning (height and age). We apply dropout regularization of 20% on the dense layer.

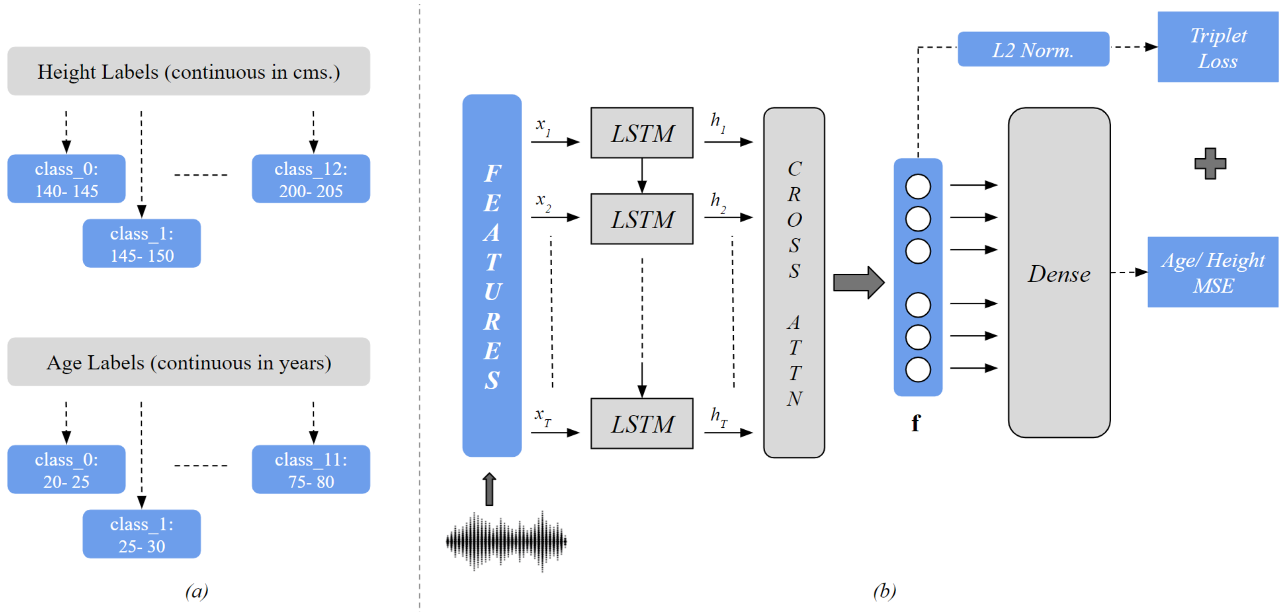

Fig. 2. Proposed Cross Attention + Triplet Loss Model for Age/ Height Estimation

Fig. 1. presents the complete architecture for this model with shapes of the output of each layer.

For extracting vector embeddings using triplet loss, we use the output of the cross attention layer, followed by an L2 normalization to produce discriminative embeddings of size 864 i.e. $(n_{frames} + n_{units})$ as shown in Fig. 2.

The final loss function for multi-task model is the weighted sum of Mean Squared Error for height and age while the final loss function for triplet loss model is the weighted sum of Mean Squared Error and Triplet Loss for height/ age estimation.

For the performance analysis of models, we use standard metrics of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which are defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \; ; \; MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{9}$$

where y_i and $\hat{y}_i$ are the actual and predicted values respectively of i-th utterance and N is the total number of utterances.

For further details and for reproducing the our work, please refer to our repository[1].

B. Quantitative Results

In the subsequent experiments, we analyze the performance of different models to show how incorporating techniques discussed in the previous sections help obtain better estimation models for height and age. First, we perform experiments with models trained only for single task i.e. individual models for height and age. We compare the performance of single-task models in three settings: model utilizing Conventional Attention (attention only across time), model utilizing our proposed Cross Attention (attention across both time and lstm units) and model utilizing Cross Attention with Gender as a binary input feature. Secondly, we compare the best performing single-task model with model trained using multi-task learning i.e. estimating height and age simultaneously using a single model. Lastly, we show how triplet loss may help to train and produce more discriminative embeddings after the cross attention layer. We also compare our best results with the previous results obtained in the literature to gain a better idea of improvements. It is also important to note that unlike most of the works in the literature in this area, we use a single model for our estimation tasks for both male and female instead of training two separate models for male and female.

In our first experiment, we compare models utilizing Conventional Attention [9], [10], proposed Cross Attention and Cross Attention with gender as a binary gender input. From Table 1, it may be seen that our proposed cross-attention mechanism significantly outperforms the conventional attention [9], [10] mechanism. Moreover, further enhancement in estimation results is obtained when gender feature is passed in a binary input to the cross attention model. It may be noted that all the models shown in Table I are trained in a single-task setting i.e. either age or height.

In our second experiment, we study the effect of training a multi-task model which simultaneously estimates height and age in a single model. As shown in Table II, multi-task learning tends to enhance the generalization ability of the model and thus, gives better results compared to single-task. RMSE of 6.95 and 6.44 cm and MAE of 5.26 and 5.15 cm for height

[1] Code Repository: https://github.com/manavkaushik/Speech-Analysis-for-Speaker-Characteristics-Estimation/tree/master

TABLE I
COMPARISON FOR PROPOSED CROSS ATTENTION

Model	***Gender***	***Height (cms)***		***Age (yrs)***	
		RMSE	MAE	RMSE	MAE
Convent. Attention	Male	7.12	5.56	8.08	5.84
	Female	6.72	5.37	9.08	6.24
Cross Attention	Male	7.04	5.45	7.96	5.60
	Female	6.64	5.25	8.92	6.16
Cross Attn. + Gender	Male	**6.98**	**5.30**	**7.90**	**5.56**
	Female	**6.50**	**5.22**	**8.70**	**6.12**

estimation for male and female respectively and RMSE of 7.81 and 8.60 years and MAE of 5.50 and 5.89 years for male and female respectively are achieved using multi-task learning. It may be noted that both the models use gender as a binary input feature.

TABLE II
MULTI-TASK AND SINGLE-TASK PERFORMANCE

Model	***Gender***	***Height (cms)***		***Age (yrs)***	
		RMSE	MAE	RMSE	MAE
Cross Attn. + Gender	Male	6.98	5.30	7.90	5.56
	Female	6.50	5.22	8.70	6.12
Multi-task Learning	Male	**6.95**	**5.26**	**7.81**	**5.50**
	Female	**6.44**	**5.15**	**8.60**	**5.89**

Lastly, we showcase how triplet loss may be employed to train more disriminative embeddings for better estimation of height and age. The triplet loss tends to effectively cluster different classes of height and age and as a result makes it easier for the final predictive dense layer to estimate the actual height and age of a person, especially in case of female speakers.

Although the results for height estimation show only slight improvement, especially for male speakers, compared to previous method and are behind the current state-of-the-art, age estimation results (for both male and female) attain the state-of-the-art performance after incorporating triplet loss as a part of the final loss function. From this difference in performance gain in age estimation compared to height estimation, it may be deduced that variation in age enables the model to produce distinctive embeddings, especially for women's age, while variation in height is not being captured as distinctively in the attention embeddings. This may be attributed to the fact that human glottal-pulse rate and vocal tract structure significantly vary with age [17]. Moreover, there are age-related vocal tract dimensional changes and concomitant decreases in all the vowel formant frequencies as people age [15] which may result in enabling triplet loss to better discriminate when embeddings are obtained on age classes. However, further investigation may help to better substantiate this difference.

It may also be noted that for age estimation task, triplet loss tends to enhance results for female characteristic estimation more than male characteristics estimation. Even in case of height estimation task, female height estimation shows more significantly improvement in results (approximately 3.5% over multi-task model) while the results for male height estimation are only marginally better and may not always be considered statistically significant. Exploring more prominent reasons for these differences for age and height estimation is a topic of further research in itself.

TABLE III
HEIGHT ESTIMATION WITH TRIPLET LOSS

Model	***Gender***	***Height (cms)***	
		RMSE	MAE
Cross Attn. + Triplet Loss	Male	6.92	5.20
	Female	6.24	**4.95**
Singh et al. [16]	Male	**6.7**	**5.0**
	Female	**6.1**	5.0
Kalluri et al. [8]	Male	6.85	-
	Female	6.29	-

TABLE IV
AGE ESTIMATION WITH TRIPLET LOSS

Model	***Gender***	***Age (years)***	
		RMSE	MAE
Cross Attn. + Triplet Loss	Male	**7.20**	**5.04**
	Female	**7.10**	**5.02**
Singh et al. [16]	Male	7.8	5.5
	Female	8.9	6.5
Kalluri et al. [8]	Male	7.60	-
	Female	8.63	-

Table IV shows that our model combining cross attention and triplet loss achieves state-of-the-art performance for age estimation on TIMIT test dataset with RMSE of 7.20 and 7.10 years and MAE of 5.04 and 5.02 years for male and female speakers respectively giving us an overall improvement of approximately 8% over other works in the literature. This proves our hypothesis that triplet loss helps to train more discriminative embeddings for enhanced estimation especially for female age.

C. Phonetic Analysis

We study which phones are important for estimation task by tracking the attention weights for each phone in the TIMIT data. We note that TIMIT contains manual phone boundaries, therefore, we can infer phone labels for each time-frame of an utterance. We accumulate the weight across all utterances to obtain an average weight for each phone.

There are a total of 60 different phones which have been used in the TIMIT dataset, and a broader analysis of attention weight distribution shows us that the highest attention weights have been assigned to Vowel phones followed by Nasal phones while the lowest attention is allocated to Stop phones. The distribution of aggregated average attention weights across different types of phones has been represented in Fig. 3 while Fig. 4 visualizes the average attention allocated to the ten most attended and ten least attended phones.

From the figures, it is clear that Vowel phones such as ’ay’, ’aw’, ’aa’, ’ae’, ’ao’, ’eh’, ’ey’, etc. tend to be most attended to while Stop phones such as ’d’, ’b’, ’p’, ’k’, etc, and some of the Fricative phones such as ’s’, ’sh’, ’f’, etc. are least attended to. Thus, it may be deduced that Vowel phones hold the highest amount of linguistic and para-linguistic information which makes them more important for the estimation task.

In order to understand the plausible reasons behind such distribution, it is important to understand that during the utterance of Stop phones, vocal folds do not vibrate, instead, their production requires a complete closure of the vocal tract. These phones are predominantly characterized by bursts of air-pressure after a short pause which may be the reason for them having the least attention weights. During the utterance of Fricatives phones, the vocal folds do not vibrate and the glottis remains open [16]. The airflow through the glottis is continuous resulting in phonetic sounds like ’s’, ’sh’, ’f’, etc. which carry little information regarding the vocal tract of a person.

On the other hand, during the utterance of Vowel and Nasal phones the vocal folds vibrate at a significant rate, resulting in periodic complete or partial closure of the glottis. This results in a pulsed airflow through the glottis, which gives these sounds their periodic nature and enables them to carry more information regarding the vocal tract of a person for better age and height estimation. Moreover, there are age-related vocal tract dimensional changes and concomitant decreases in all the vowel formant frequencies as people age [15] which may result in enabling the attention models to predict age more accurately, especially while attending more to vowel phones.

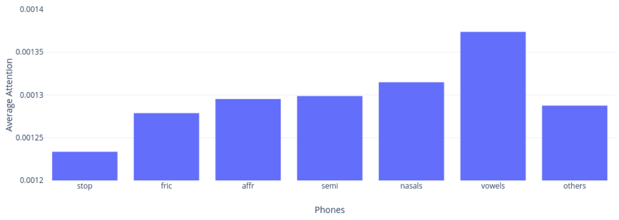

Fig. 3. Attention distribution among different types of phones

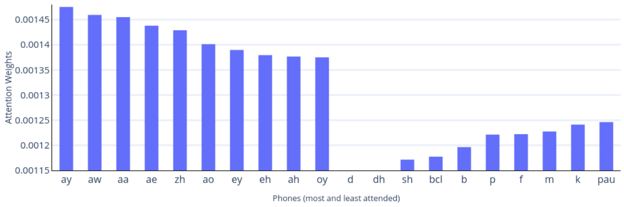

Fig. 4. 10 Highest & 10 Least Attended Phones respectively

VI. CONCLUSIONS

We have proposed a cross-attention approach for the task of joint speaker height and age estimation. The proposed approach not only performed soft-attention across time-frames but also performed soft-attention across hidden units which produces more informative context vector. Experimental results on TIMIT data show that our proposed approach outperforms conventional attention mechanism and gave consistently better results when tested in multi-task setting by gaining better generalization for effective estimation. Furthermore, using vector embeddings trained from triplet loss further enhances the age and height estimation task and achieves state-of-the-art performance for age estimation on TIMIT dataset which may be attributed to the fact that vocal tract length and glottal-pulse rate vary with variation with age and height of a person, thus, making it possible to for triplet loss to capture these variations in vocal tract by discriminating amongst the quantized classes of age. However, we acknowledge that further research and investigation may be required to consolidate reasons for a difference in improvement of age estimation and improvement of height estimation and also for variations in improvement in male and female counterpart after employing triplet loss.

Finally, by tracking attention weights across time-frames, we found that Vowel phones are most important while Stop phones and Fricatives have been least attended by the attention model for speaker physical characteristics estimation.

At the same time, we do note that our height performance results with triplet loss are borderline in terms of the statistical significance and thus, are a matter of further investigation. Moreover, there is further scope of improvement, especially for height estimation, which may be brought about by utilizing age and height information that is available through several large public corpora. We plan to further investigate in this area by using techniques such as unsupervised and self-supervised learning.

REFERENCES

[1] Tanner, D.C. and Tanner, M.E., “Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection,” *Lawyers & Judges Publishing Company* , 2004. [Online]. Avail-able: https://books.google.com.sg/books?id=u39sykyx2zwC.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, “Paralinguistics in speech and language - state-of-the-art and the challenge,” *Computer Speech andLanguage, Special Issue on Paralinguistics in Naturalistic Speechand Language,* 2013.

[3] K. A. Williams and J. Hansen, “Speaker height estimation combining gmm and linear regression subsystems,” in *Proc. of ICASSP,* 2013, pp. 7552–7556.

[4] A. H. Poorjam, M. H. Bahari, V. Vasilakakiset al., “Height estimation from speech signals using i-vectors and least-squares support vector regression,” in *Proc. of ICTSP,* IEEE, 2015, pp. 1–5.

[5] D. Mahmoodi, H. Marvi, M. Taghizadeh, A. Soleimani, F. Razzazi, and M. Mahmoodi, “Age estimation based on speech features and support vector machine,” in *Proc. of CEEC,* IEEE,2011, pp. 60–64.

[6] T. Bocklet, A. Maier, and E. Noth, “Age determination of children in preschool and primary school age with gmm-based super vectors and support vector machines/regression,” in *Proc. of International Conference on Text, Speech and Dialogue,* Springer, 2008, pp. 253–260.

[7] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, “End-to-end deep neural network age estimation,” in *Proc. of INTERSPEECH,* 2018, pp. 277–281.

[8] 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 *Proc. of ICASSP,* IEEE, 2019, pp. 6580–6584.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of ICML,* 2015, pp. 2048–2057.

[10] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end models for small-footprint keyword spotting," in *arXiv preprint arXiv:1803.10916,* 2018.

[11] Cheng, Jianpeng Dong, Li Lapata, Mirella. "Long Short-Term Memory-Networks for Machine Reading," 551-561. 10.18653/v1/D16-1053, 2016.

[12] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of machine learning research,* vol. 10, no. 2, 2009.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2015, pp. 815–823.

[14] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737,* 2017.

[15] S. A. Xue and G. J. Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production," *ASHA,* 2003.

[16] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *Proc. of IWBF,* IEEE, 2016, pp. 1–6.

[17] P. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *The Journal of the Acoustical Society of America,* vol. 118, no. 5, pp. 3177–3186, 2005.

[18] M. H. Bahari, M. McLaren, D. Van Leeuwenet al., "Age estimation from telephone speech using i-vectors," in *Proc. of INTERSPEECH,* Portland, USA, 2012.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *arXiv preprint arXiv:1409.0473,* 2014.

[20] V. Mnih, N. Heess, A. Graveset al., "Recurrent models of visual attention," in *Advances in neural information processing systems,* 2014, pp. 2204–2212.

[21] V. Mnih, N. Heess, A. Graveset al., "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755,* 2014.

[22] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition,* vol. 48, no. 10, pp. 2993–3003, 2015.

[23] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium,* 1993.

[24] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779,* 2019.

[25] R. Zazo, P. S. 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. 22524–22530, 2018.

[26] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. of ICML,* 2013, pp. 1310–1318.