Online-Journals.org

iJOE International Journal of
Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 10 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i10.39061

PAPER

End-to-End Speaker Profiling Using 1D CNN Architectures and Filter Bank Initialization

Umniah Hameed Jaid[1,2](✉),
Alia Karim AbdulHassan[2]

[1]University of Baghdad, Baghdad, Iraq

[2]University of Technology, Baghdad, Iraq

umniah.h@sc.uobaghdad.edu.iq

ABSTRACT

The automatic estimation of speaker characteristics, such as height, age, and gender, has various applications in forensics, surveillance, customer service, and many human-robot interaction applications. These applications are often required to produce a response promptly. This work proposes a novel approach to speaker profiling by combining filter bank initializations, such as continuous wavelets and gammatone filter banks, with one-dimensional (1D) convolutional neural networks (CNN) and residual blocks. The proposed end-to-end model goes from the raw waveform to an estimated height, age, and gender of the speaker by learning speaker representation directly from the audio signal without relying on handcrafted and pre-computed acoustic features. The conducted experiments on the TIMIT dataset show that the proposed approach outperforms many previous studies on speaker profiling with a mean absolute error (MAE) of 5.18 and 4.91 cm in height estimation and MAE of 5.36 and 6.07 years in age estimation for males and females, respectively, and achieving an accuracy of 99.98% in gender prediction.

KEYWORDS

age estimation, height estimation, gender detection, gammatone filter bank, wavelet filter bank

1 INTRODUCTION

A speech signal conveys more than just linguistic information; it can indicate the identity of the speaker as well as cues about their age, gender, ethnicity, and emotions [1]. Voice characteristics can provide indications of a speaker's age; for instance, younger speakers tend to have a higher speech rate [2], while the fundamental frequency decreases with age [3]. Additionally, the fundamental frequency can help differentiate between male and female speakers, as males usually have lower frequencies. Height estimation can be linked to the assumption that the vocal tract length is directly proportional to the speaker's height [4]. The current research suggests the ability to estimate physical parameters such as height [1, 2], age, and gender from speech only [3, 4]. The extraction of these parameters can find applications in various areas, such as forensics [5], targeted marketing, customer services [2, 6],

Jaid, U.H., AbdulHassan, A.K. (2023). End-to-End Speaker Profiling Using 1D CNN Architectures and Filter Bank Initialization. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(10), pp. 65–81. https://doi.org/10.3991/ijoe.v19i10.39061

Article submitted 2023-02-22. Resubmitted 2023-04-27. Final acceptance 2023-04-27. Final version published as submitted by the authors.

interactive healthcare systems, and human-robot interaction [7]. These applications often require accurate responses in a timely fashion. A system aimed at predicting a speaker's profile should be able to provide an approximate speaker profile from short speech utterances of varying lengths that are produced in a noisy environment, and thus the feature extraction and processing methods should be robust and minimal.

Speaker profiling is considered a challenging task due to the overlap of many factors that affect the human voice, such as emotional state, health, weight, and the context of the speech. However, one of the most challenging aspects of speaker profiling is distinguishing textual content from physical traits [3].

Typically, speaker profiling involves three stages: data gathering and preprocessing, feature extraction and selection, followed by estimation and prediction of physical traits. Of these, feature extraction and selection are the most important, where raw speech signals are often converted to time-frequency representations to capture patterns that appear in the voice. For instance, features such as mel-frequency cepstral coefficients (MFCC) [8], linear predictive coding (LPC) [9], and formant frequencies [10] are extensively used in speaker profiling. Other studies adopt statistical features, such as mean, median, and percentiles, for speech utterance representation [3, 9, 11]. Another approach involves Gaussian mixture models—universal background model (GMM-UBM) and i-vectors for modeling speech utterances [12–14].

Deep-learning (DL) approaches are powerful in extracting and learning complex patterns from data. The deep architecture of DL models, which comprises multiple hidden layers, enables them to perform better than many machine-learning algorithms in speaker profiling [15].

Deep learning includes various types of neural network architectures employed for different speaker-profiling tasks, for instance, Long short-term memory (LSTM) recurrent neural networks joined with different features, including MFCC, pitch, and normalized cross-correlation, function to perform age estimation [16].

Kalluri et al. [17] explored the use of deep neural networks (DNN) to jointly estimate height and age from a speech by utilizing a support vector regression (SVR) model trained with GMM mean supervectors for DNN initialization. Similarly, Kaushik et al. [18] investigated the use of DNN for height and age estimation by proposing an LSTM architecture with an attention mechanism. LSTMs were also utilized in conjunction with CNNs by [19] as speech encoders in a semi-supervised manner where the encoded speech is then fed into a two-layer NN for fine-tuning and joint estimation of height, age, and gender. The use of x-vectors and d-vectors is explored in combination with transfer learning in the joint estimation of age and gender [20, 21], as the work addresses the matter of limited training data by employing transfer learning from networks pre-trained for different tasks other than speaker profiling, utilizing well-known speech recognition datasets, such as Librispeech [22] and common voice datasets [23]. The use of transfer learning showed an improvement in terms of MAE and root mean squared error (RMSE).

However, the majority of previous work in speaker profiling has employed deep learning on hand-crafted features [17] or relied on complex structures involving millions of parameters [19, 24]. This approach has several drawbacks, including the high cost of feature extraction and dependency on manually selected features, which can lead to poor performance if they are not suitable for the task at hand. Additionally, the use of complex models with a large number of parameters necessitates a substantial amount of training data [24]. These methods can be time consuming and may not be efficient when dealing with large amounts of data, particularly in real-time applications.

The processing of raw waveforms is a widely studied topic in various facets of speech recognition [25], speaker identification [26], and other sound-related tasks [27]. Many studies in this area make use of adapted versions of CNNs to process the signal. For example, the authors of [25] proposed SincNet, a specific convolutional layer that replaces conventional convolutional layer weights with acoustic filter parameters, such as band-pass filters.

Another study builds on the work of [25] by replacing SincNet rectangular band-pass filters with Gaussian and gammatone filters [28]. A continuous wavelet convolutional layer is proposed in [29] to replace the first convolutional layer, where two parameters, scale and translation, are learned.

This work proposes a compact, true end-to-end speaker-profiling system that aims to jointly estimate the age, height, and gender of an unknown speaker from short utterances, using raw waveforms directly with minimal pre-processing. Moreover, unlike previous studies that handle raw waveforms, instead of replacing the first layer with a filter where only the filter parameters are learned, our study proposes to initialize the first layer with a filter bank and allow the network to update the values of the filter during training. This, in turn, provides a balance between handcrafted feature extraction and representation learning by a CNN. Using this approach enables efficient creation of a personalized filter bank, which is specifically optimized to extract speaker-related characteristics from the signal. Two initialization methods are employed in the first layer of the CNN; namely, continuous wavelet transforms and the gammatone filter bank.

The proposed system is able to outperform existing systems employing DNN with handcrafted time-consuming feature extraction and processing methods. Contributions of this work can be summarized as performing multi-task speaker profiling for age, height, and gender from raw speech and presenting a novel wavelet filter-bank initialization method for CNN with residual blocks.

This paper is organized as follows: Section 2 presents the building blocks of the proposed end-to-end model and the proposed approach to initialize the first layer of the CNN with gammatone and wavelet transforms. Section 3 presents the experimental setup, including the dataset used and the evaluation metrics used to evaluate the model performance. Section 4 analyzes various aspects of speaker profiling with the proposed model, such as the effect of speech duration on the performance of the model, the effects of single-task and multi-task predictions, and the effect of gender information on the accuracy of height and age predictions. Our conclusions from this work are presented in the last section.

2 PROPOSED MODEL

Although DNNs have been successfully applied to many speaker-profiling tasks, they typically rely on handcrafted features and complex models that involve a large number of parameters. The proposed end-to-end model aims to bypass the stages of pre-processing and feature extraction and learn a discriminative representation of the speaker directly from the audio signal with minimum pre-processing effort and a compact architecture with few parameters. The proposed model consists of three main components: 1D convolutional layers, residual blocks, and filter-bank initialization. Subsections 2.1 and 2.2 provide an overview of each component, with Subsection 2.3 describing the proposed model in detail.

2.1 CNN

A CNN is a class of deep-learning networks that is capable of extracting features from raw data, usually an image. A CNN consists of several convolutional layers combined with pooling layers and fully connected layers [30]. A 1D CNN is a CNN with an array representing the audio waveform as input. A convolutional layer consists of a number of kernels or filters that convolves through the input to produce a feature map as in Eq. (1):

$$f\_map_k = x \times w_k^n + b^n \tag{1}$$

where *f_map*$_k$ is the k-th resulting feature map, *x* is the input signal, w_k^n is the weight of the k-th kernel of the n-th layer, and b^n is the bias term of the n-th layer. Afterwards, an activation function is applied to the resulting feature maps to apply nonlinear transformations. The process of convolving the input with k kernels results in a large number of feature maps. The pooling layer is then employed, not only for downsampling but also for reducing the spatial dimensions and computational complexity of the resulting feature maps, while preserving their most important features. Various methods, such as average pooling, max pooling, or other statistical measures, can be used to reduce the size of the resulting feature maps.

For a model to be able to learn appropriate and relevant features from the input signal, several convolutional layers are stacked to arrive at a final feature map that is passed to a fully connected layer for the final prediction of the network. For the regression task, a rectified linear unit (ReLU) activation function is used to obtain the final prediction of the network as in Eq. (2):

$$f(x) = \max(0, x) \tag{2}$$

CNNs are designed to perform feature extraction on raw data; however, given the complex nature of speech utterances, deeper and more complex models are needed to handle raw signal inputs, which, in turn, can lead to the vanishing gradient problem [16], as the gradient of the training error gets closer to zero, restricting the learning process and causing the model's accuracy to degrade. To alleviate this problem and allow the model to learn the hidden structures of the speech signal, residual blocks are included in the model instead of traditional convolutional layers.

2.2 Residual blocks

In traditional neural networks, the output of each layer is directly fed to the subsequent layer. However, residual networks (ResNets) introduce a different approach by incorporating skip connections. In a Res-Net, the output of a layer is fed not only to the next layer but also to layers n steps ahead. These skip connections help improve gradient flow during backpropagation, allowing for deeper network architectures without the risk of vanishing gradients.

Residual blocks, which are also known as skip-connection blocks, are the building blocks of ResNets [31]. They learn residual functions with reference to the layer input. A residual block is shown in Figure 1.

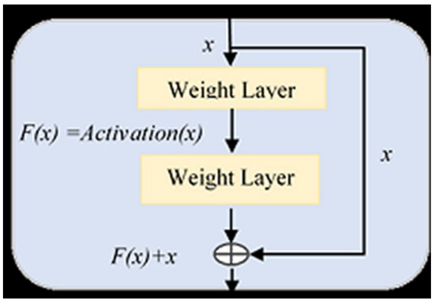


Fig. 1. General structure of a residual block

2.3 Filter banks

This section briefly describes two types of filter banks that have been utilized in the proposed model to initialize the first layer of the network: gammatone filter banks and wavelet filter banks. These filter banks play a crucial role in extracting relevant features from speech signals by focusing on specific frequency bands, thereby improving the performance of the system.

Gammatone filter. A gammatone filter bank resembles the basilar membrane motion in the human auditory system. It consists of non-linearly spaced band-pass filters with increasing bandwidth. According to [32], the impulse response of gammatone is given as Eq. (3):

$$\gamma(t) = at^{(p-1)}e^{-2\pi\beta t}\cos(2\pi f_c t + \Phi) \tag{3}$$

where f_c is the center frequency, β represents the filter bandwidth, a is the amplitude, p is the order of the filter which takes values 2 to 8, and Φ is the phase shift. The bandwidth and central frequency parameters are converted to equivalent rectangular bandwidth (ERB) [33] using Eq. (4):

$$ERB(fc) = 24.7 + fc.\,9.256 \tag{4}$$

To obtain a gammatone filter bank, a process similar to that of [27] is followed. Accordingly, a range of center frequencies is usually predetermined between the lowest frequency in the human hearing range and the sampling rate divided by 2, which is the Nyquist frequency. Then, the central frequencies are placed equally on an ERB scale, such that the next frequency is obtained by starting with the lower frequency converted to ERB with (4), increasing by one, and then converting back to the frequency domain.

For the gammatone filter bank, the central frequency is set in the range between 62 Hz and 8000 Hz, which is the Nyquist frequency for a sampling rate of 16,000 Hz. The filter response is then obtained for each center frequency, resulting in 64 filters which are then used to initialize the kernels of a convolutional layer.

Wavelet filter. Wavelet analysis is a powerful tool for signal analysis since it allows the frequency and time localization of a signal to be determined.

The fundamental concept behind wavelets is that any speech signal can be analyzed with a mother wavelet using different scales, as illustrated in Figure 2.

Continuous wavelet transforms (CWT) allow for arbitrary time-frequency scales limited only by the sampling rate. Higher scales in the time domain correspond with small frequencies and vice versa. In general, a CWT is described by:

$$X_{\omega(s,u)} = \frac{1}{|s|^{1/2}}\int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-u}{s}\right)dt \tag{5}$$

where Ψ is the mother wavelet function, s is the scale factor, and u is the translation factor. Given a signal X of length N, and a wavelet function Ψ, for example, the Morlet wavelet, the continuous wavelet transform of the signal X is defined as the convolution of X with scaled and translated wavelet Ψ and the convolution should be repeated n times for each scale [34].

A filter bank can be created by utilizing a collection of scaled and shifted wavelets as filters to analyze a signal's behavior across specific frequency bands. In this configuration, according to the input size and the required length of the filter, a Morlet wavelet transform matrix is calculated by first estimating the required scales based on the required number of filters. The transform at each scale is calculated by Eq. (6):

$$\psi = \pi^{1/4} e^{i\omega_0 t}\ e^{-t^2/2} \tag{6}$$

where ω_0 is the non-dimensional frequency constant, defaulted to 6 [34, 35], and t is a time parameter. The wavelet at each scale is then represented as a filter, and these filters are used to initialize the filters of the first layer of the network.

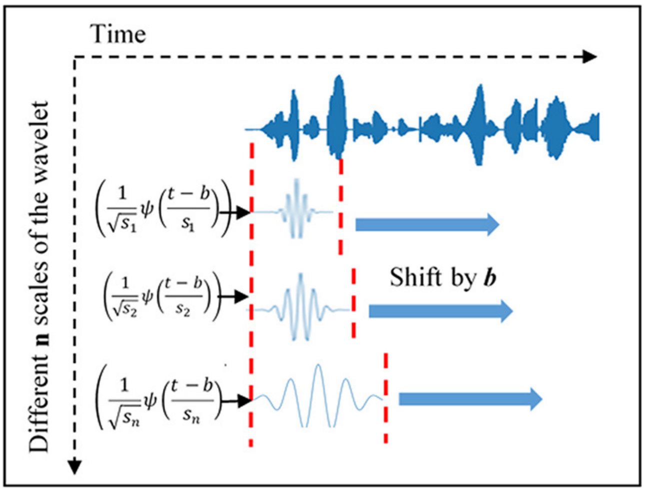

Time
Different **n** scales of the wavelet
$\left(\frac{1}{\sqrt{s_1}}\psi\left(\frac{t-b}{s_1}\right)\right)$
$\left(\frac{1}{\sqrt{s_2}}\psi\left(\frac{t-b}{s_2}\right)\right)$
$\left(\frac{1}{\sqrt{s_n}}\psi\left(\frac{t-b}{s_n}\right)\right)$
Shift by ***b***

Fig. 2. The process of convolving the signal with multiple continuous wavelet filters with different scales

When using a wavelet transform to analyze a signal, the choice of the number of scales to use can have a significant impact on the results. One way of choosing the scales is to use a random set of scales; however, according to [34], a more appropriate choice is to write the scales as fractional powers of two, as in Eq. (7):

$$s_j = s_0 2^{j\delta j},\ j = 0, 1, \ldots, J \tag{7}$$

where s_j is the jth scale, s_0 is the minimum scale, δj is the resolution of the scale, and J is the largest scale that is obtained using Eq. (8):

$$J = \frac{1}{\delta j}\log_2 N\delta t/s_0 \tag{8}$$

where N is the signal length, and δt is the signal time steps. The minimum scale is chosen such that the Fourier period is around $2\delta t$.

2.4 Proposed architecture

The proposed architecture, shown in Figure 3, consists of a 1D convolutional layer for feature extraction, three residual blocks, and two fully connected layers. Each residual block consists of three 1D convolutional layers and one max pooling layer. The first layer used to process raw audio signals employs a large receptive field to have a more global view of the audio, while the subsequent layers have a shorter filter size of two to aid the CNNs in learning hierarchical representations using its depth. A dropout layer is attached after each residual block to reduce the effects of overfitting. Two fully connected layers are added after the convolution layers as output layers, where one output layer consists of two neurons for the regression task of height and age estimation that employs a ReLU activation function and a mean absolute error (MAE) loss function. The other output layer uses a sigmoid activation and binary cross entropy as a loss function for gender classification into male or female.

To enhance the feature-extraction capabilities of the model, the characteristics of convolutional layers are exploited by incorporating auditory filter banks as initialization methods for the first layer of the proposed architecture. One filter bank is the gammatone filter bank inspired by [27]. The other filter bank is inspired by a multiscale continuous wavelet transform. The use of such an initialization method can be considered complementary to the role of the CNN layers as feature extractors and aid as a compromise between handcrafted features and representation learning.

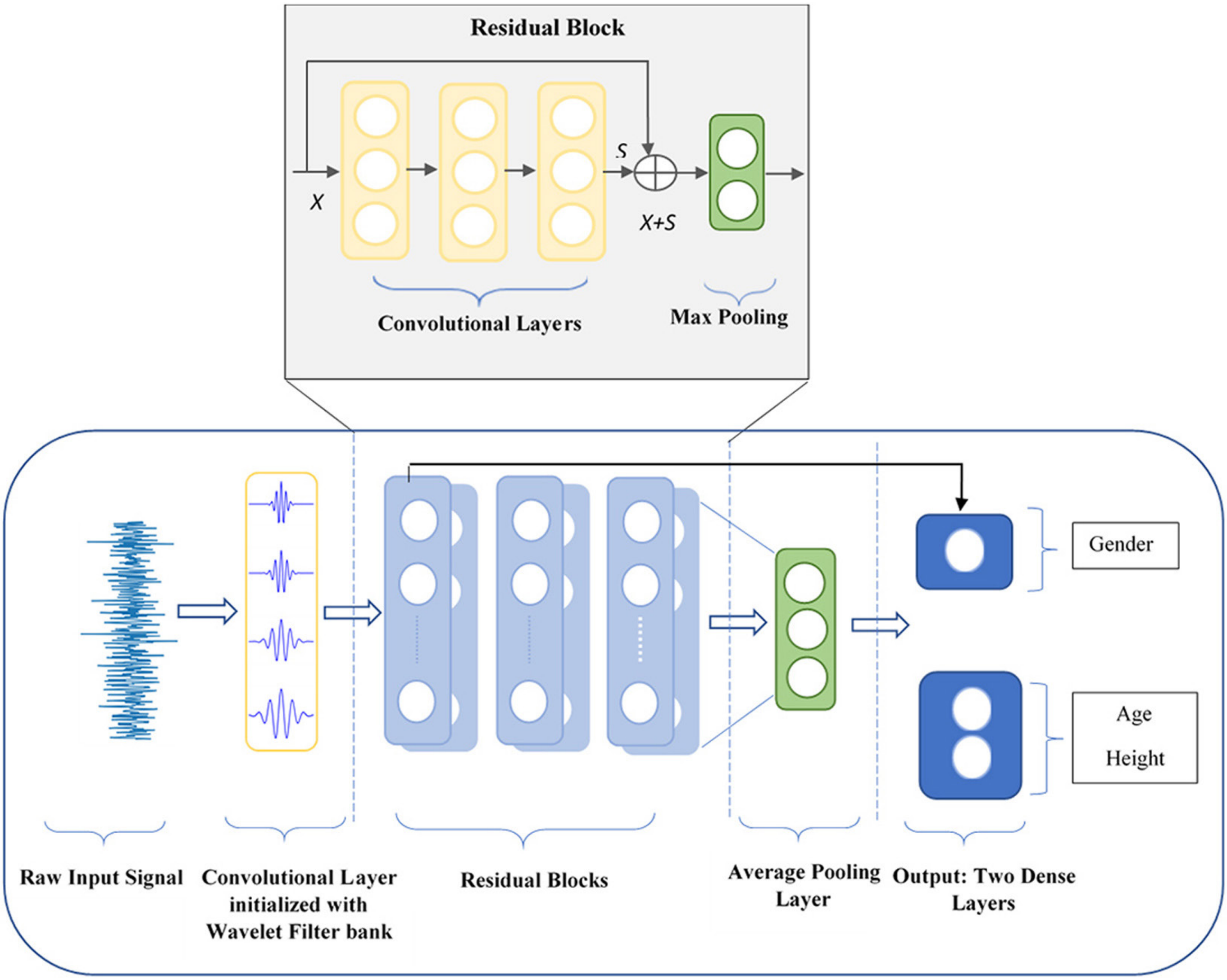


Fig. 3. Proposed Architecture for joint estimation of height, age, and gender from speech, featuring wavelets as an example of a filter bank for initialization

3 EXPERIMENTAL SETUP

For the joint height, age, and gender estimation experiments, the TIMIT dataset [36] is used with the standard train and test split with 10% of the training data set as validation data. The evaluation metrics used are MAE, and RMSE defined as:

$$MAE = \frac{\Sigma\left|y_i - x_i\right|}{n} \tag{9}$$

$$RMSE = \sqrt{\frac{\Sigma(y_i - x_i)^2}{n}} \tag{10}$$

where y_i is the predicted value, x_i is the target value, and n is the number of observations and accuracy for the gender classification:

$$\text{Accuracy} = \frac{No.\ of\ \ Correct\ \ Predictions}{Total\ No.\ of\ \ Predictions} \tag{11}$$

Since the number of convolutional layers plays a key role in the performance and complexity of the network, the number of convolutional layers for the proposed network was determined by experiment. Table 1 shows the architecture used in the following experiments, and Table 2 shows the structure of each residual block. For each experiment, the network was trained up to 200 epochs with batch sizes of 64 samples.

Table 1. DNN architecture employed in all experiments

Layer Type	No. of Filters	Kernel Size/Strides
Conv1D, ReLU	64 filter	100/100
Residual Block	32	3/1
Residual Block	64	3/1
Residual Block	64	3/1
Average Pooling 1D	–	3/3
Dense, ReLU	2	–
Dense, Sigmoid	1	–
Total No. of Parameters		77,603

Table 2. Design of residual block

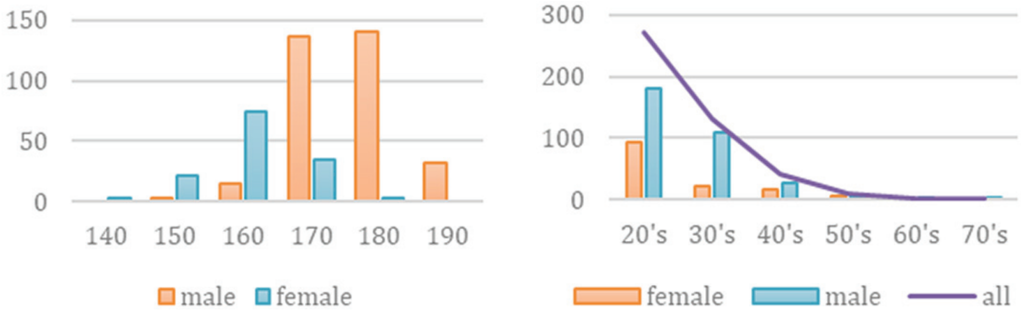
Layer Type	Kernel Size/Strides
Conv1D	1/1
Conv1D, ReLU	3/1
Conv1D	3/1
Add Layer	–
Max Pooling	2/2

As described in Section 2, the first layer of the network is the equivalent of a feature extractor from raw audio using a gammatone filter bank or a CWT filter bank.

For the gammatone filter bank, 64 filters of length 100 and no overlap are initialized with a gammatone impulse response as in Eq. (3), where each filter represents the response at some center frequency in the range of 100 Hz to 8000 Hz. After the initialization, the filters are set to be trainable to allow the network to learn new filters. As for the CWT initialization, the number of filters is determined as the optimal number of scales required to process the input signal. Each filter represents a scaled version of the mother wavelet. In this setting, 40 filters of length 100 and no overlap are chosen to initialize the network.

3.1 Datasets

In this work, the TIMIT dataset [36] is employed to estimate the height, age, and gender of unknown speakers. The TIMIT dataset consists of 630 speakers (192 female, 438 male), and for each speaker, there are 10 audio recordings. The data is split into 462 speakers for training and validation, and 168 speakers for testing. Figure 4 shows the distribution of male and female height and age in the TIMIT dataset. As can be seen, the majority of speakers are male speakers in their 20’s and 30’s, with heights of 170–180 cm, unlike female speakers, whose heights are concentrated in the range of 160–170, and almost no data for other height ranges. The figures also show that only a few speakers are between the ages of 40 and above.



(a) Height Distribution

(b) Age Distribution

Fig. 4. The distribution for male and female speakers in the TIMIT dataset across different height and age groups

4 RESULTS AND DISCUSSION

In this section, the results of several experiments are discussed to evaluate the proposed model and the effect of duration on the prediction task, as well as the effect of multi-task vs. single-task prediction. In addition, we evaluate the effect of gender information on the prediction of height and age.

4.1 Duration analysis

To determine the minimum speech duration required for accurate profiling, we evaluated the performance of the system on TIMIT datasets with speech durations ranging from 1 s to 5 s. To evaluate longer speech durations, shorter segments were extended by appending portions of the original recording. Random initialization was used, and gender was not considered in the estimation accuracy. The duration analysis, as shown in Figure 5, indicates that the system can predict both height and age with a high degree of accuracy on durations as short as 1 second, achieving MAEs of 5.7 and 5.3, respectively.

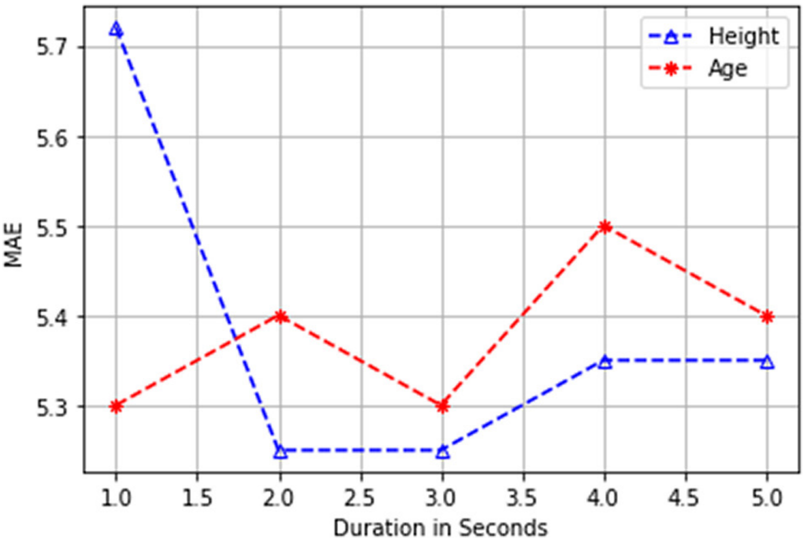


Fig. 5. MAE of height and age prediction using different durations of speech on the TIMIT dataset

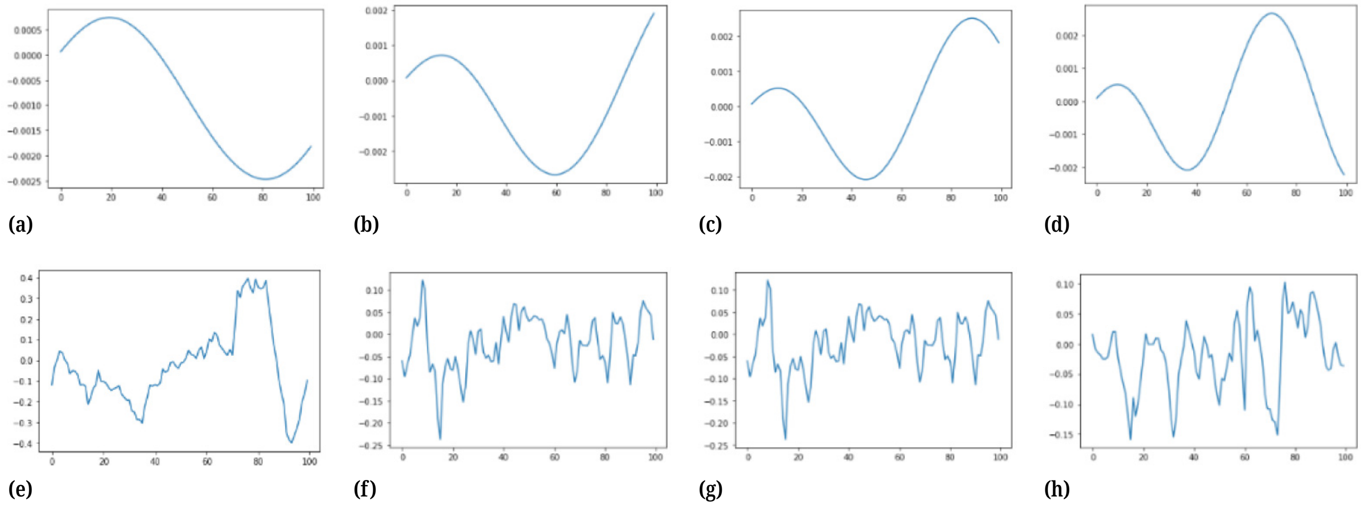
However, the best performance is achieved with recordings lasting between 2 and 3 seconds, which may be due to the original length of many recordings. It is worth noting that even with a duration as short as 1 second, the system is capable of achieving performance comparable to longer durations, indicating the efficiency and robustness of the system. Furthermore, the observation that both height and age predictions follow the same pattern with different durations suggests that the system can capture the relevant information for both features within the same timeframe. This implies that the minimum duration required for the profiling task may not necessarily be long, and comparable performance can be achieved with shorter speech segments. To ensure consistency in speech duration and improve the reliability of the comparison across different audio samples, raw audio files longer than 3 seconds were truncated, and those shorter than 3 seconds were duplicated and truncated to 3 seconds. This resulted in an audio length of 48,000 bits for a sampling rate of 16 kHz.

4.2 DNN initialization

In this section, the effect of training on the filters used to initialize the network is analyzed. Figures 6 and 7 show four randomly selected filters used to initialize the network and the same four filters learned after training for gammatone and CWT, respectively. Each filter corresponds to a CNN filter in the first layer of the network and is defined on a scale of 0 to 100, indicating its size.

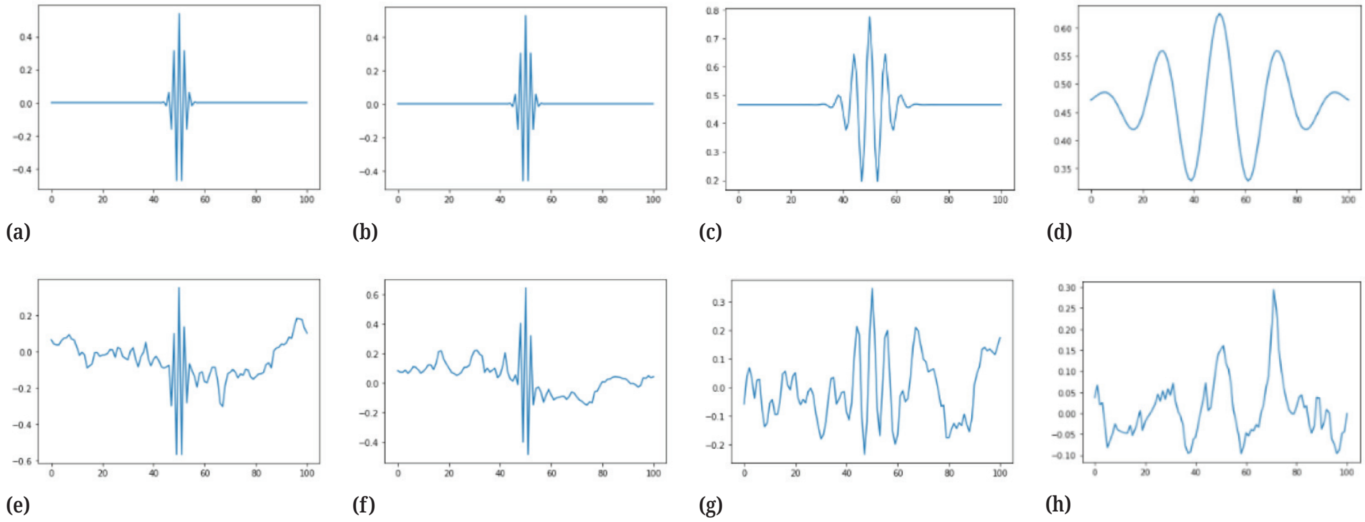
Figure 6 shows the effect of training on the gammatone filters used to initialize the first layer of the network. It is obvious that the filters with larger scales (i.e., higher frequencies) maintain the original shape after training but lose smoothness; however, the filters with smaller scales (i.e., lower frequencies) lose the original shape.

Similarly, for wavelet filters in Figure 7, the filters with the smaller scales (i.e., higher frequencies) are least affected by the training, whereas the filters with the larger scales (i.e., lower frequencies) are more distorted. This indicates that the network has learned to focus on the fine-grained details in the input data and has disregarded the coarser features represented by the larger-scale wavelets, and that the fine details that appear in the higher frequencies are important to the network and contribute more information than those of lower frequencies.


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. Visualization of randomly selected filters of the gammatone initialized layer, where (a), (b), (c), and (d) are filters before the training process, while (e), (f), (g), and (h) shows the same filters after training


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7. Visualization of randomly selected filters of the wavelet-initialized layer, where (a), (b), (c), and (d) are filters before the training process, while (e), (f), (g), and (h) show the same filters after training

To evaluate the effectiveness of the initialization scheme on the performance of the system, several initialization methods are tested, including initializing with ones, randomly, and with He Uniform initialization [37]. Figure 8 shows the effect of the initialization on the performance of the network. As can be seen, the best-performing approach for height and gender is the gammatone filter bank, followed by the CWT approach, while CWT outperforms other initialization methods in age estimation compared to the performance of other initialization methods.

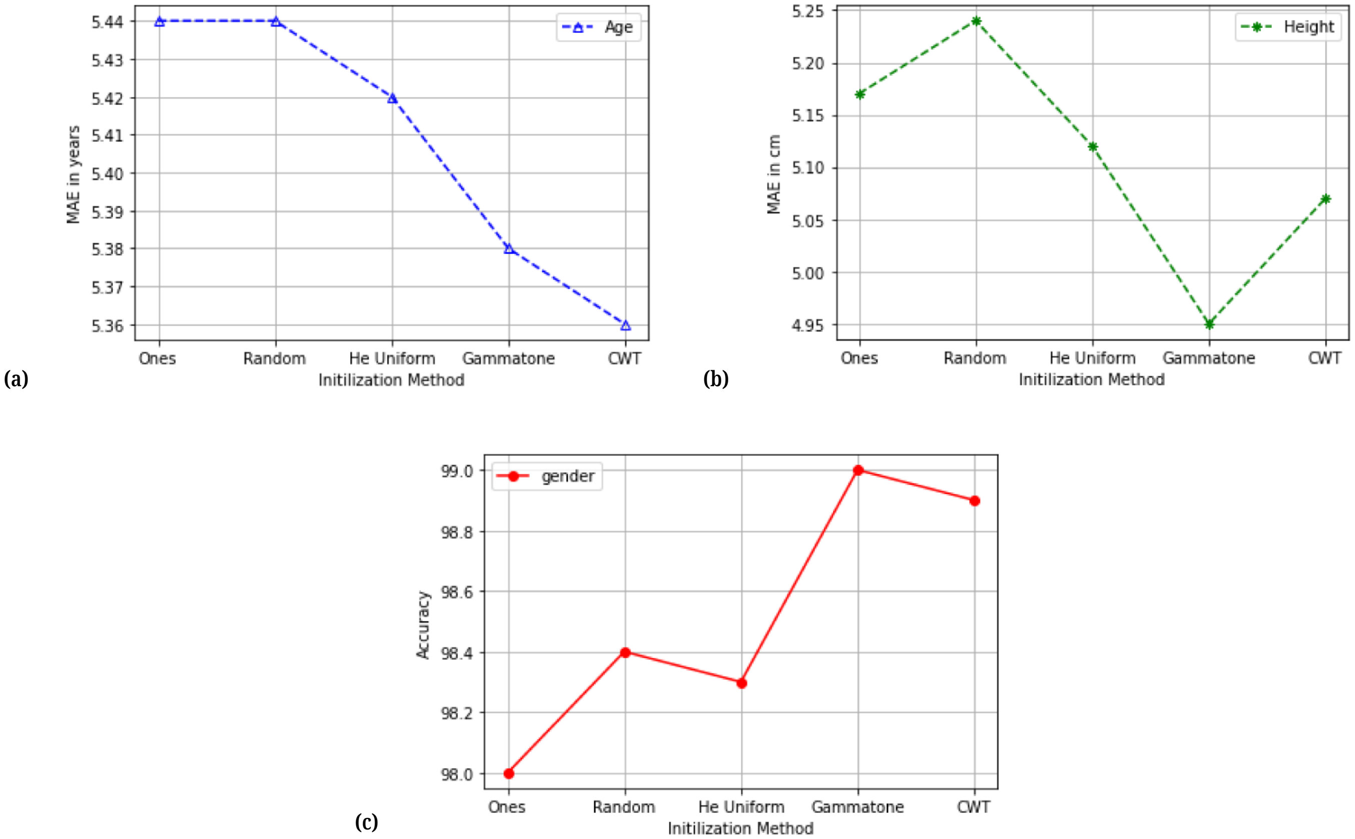


Fig. 8. Effect of different initialization methods on (a) age, (b) height, (c) gender

Moreover, the choice of wavelet transform seems to have an effect on the results as well, where three types of wavelet functions are compared. As shown in Table 3, Morlet wavelet performs better compared with other types of wavelets, such as the derivative of Gaussian (DOG) wavelet and Paul wavelet. Figure 9 illustrates the performance difference in MAE for height and age for the three types of wavelets.

Table 3. Results of different wavelet initializations on the MAE and RMSE of joint height, age and gender prediction

Initialization Method	Height				Age				Gender
	RMSE		MAE		RMSE		MAE		
	Male	Female	Male	Female	Male	Female	Male	Female	
Morlet	7.12	6.35	5.4	5.08	8.24	9.57	5.47	6.04	98.6
Paul	7	6.58	5.52	5.1	8.5	9.97	5.4	6.1	98.2
DOG	7.15	6.49	5.58	5.13	8.18	9.5	5.54	6.07	98.5

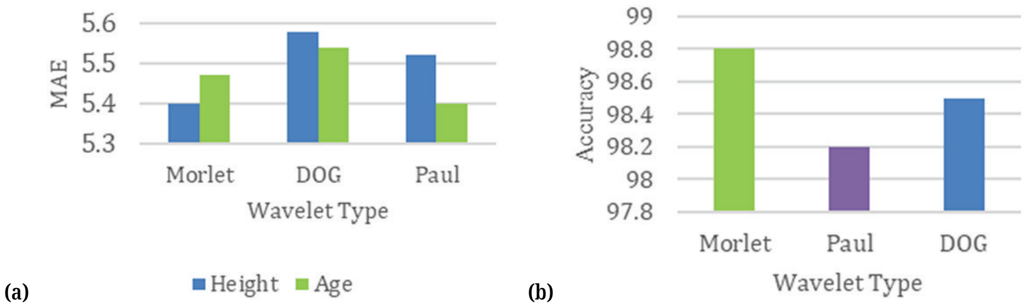


Fig. 9. Performance of different wavelet types on the estimation of (a) Height and Age, and prediction of (b) Gender

4.3 Single-task vs. multi-task

To evaluate the effect of multi-task learning and the effect of each trait on the learning outcome, several combinations are tested using the same experimental settings; all the experiments are performed on the TIMIT dataset with the same hyperparameters using the gammatone filterbanks for initialization. The experiments include single-task learning of height and age. Additionally, to evaluate the effect of gender information on the learning process, two models are trained independently on male and female data separately. The different combinations include the following: age and gender; height and gender; height and age; and height, age, and gender.

In Tables 4 and 5, the experimental results show a clear advantage of multi-task learning over single-task. However, gender information did not seem to improve the results for either height or age. The degradation in performance, which appears for the gender-based learning in both tasks of height and age estimation, can be largely attributed to the data imbalance found in the dataset. Furthermore, the training the model for height-gender and age-gender did not improve the results over single-task learning (ST) of height and age. Figure 10 presents the difference in performance between gender-based training vs. training using data from both genders.

Table 4. Results of height estimation in single and multi-task setting on the TIMIT test set

Task	RMSE		MAE	
	Male	Female	Male	Female
Height ST	7.5	6.78	5.8	5.4
Height ST (Gender-based)	7.83	7.43	6.0	5.93
Height & Gender	7.66	7.27	5.93	5.8
Height, Age & Gender	7.17	6.26	5.58	4.95

Table 5. Results of age estimation in single and multi-task setting on the TIMIT test set

Task	RMSE		MAE	
	Male	Female	Male	Female
Age ST	8.44	9.73	5.42	6.16
Age ST (Gender-based)	8.6	10.2	5.39	6.6
Age & Gender	8.57	10.0	5.5	6.2
Height, Age & Gender	8.38	9.8	5.38	6.05

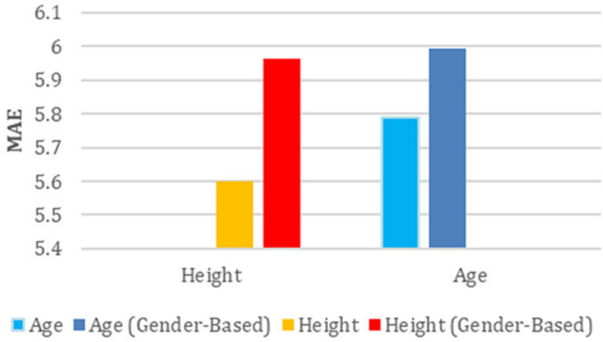


Fig. 10. The effect of gender information on the estimation of height and age

4.4 Speaker-level predictions

Since the TIMIT dataset contains 10 utterances for each speaker, the effect of prediction is evaluated at the speaker level rather than the utterance level, which can be done by aggregating the results of each speaker over their utterances and obtaining the mean value of the prediction. The results show an improvement in gender prediction and height prediction for male and female speakers. However, age prediction shows no improvement. Table 6 shows the comparison (RMSE and MAE) of our results with the previous methods for multi-task height and age and gender prediction. The results show an improvement in male and female height prediction and increased accuracy in gender detection.

Table 6. Comparison of results for height, age, and gender on TIMIT test

Study	Height				Age				Gender Accuracy
	RMSE		MAE		RMSE		MAE		
	Male	Female	Male	Female	Male	Female	Male	Female	
[18]	6.95	6.44	5.26	5.15	7.81	8.60	5.50	5.89	–
[19]	7.5	6.5	5.8	5.1	6.8	7.4	4.8	5.0	99.1
[17]	6.85	6.29	–	–	7.60	8.63	–	–	–
Proposed	7.17	6.26	5.58	4.95	8.38	9.8	5.35	6.07	99
Proposed-Aggregated	6.78	6.27	5.18	4.91	8.41	9.88	5.36	6.07	99.98

5 CONCLUSIONS

In this work, the possibility of using raw waveform for the prediction of height, age, and gender of an unknown speaker was examined. In particular, a novel approach for speaker profiling from raw audio signals was proposed, utilizing convolutional layers for feature extraction from raw input, and integrating acoustic filter banks, such as wavelets and gammatone filter banks, with the convolution process. The model was trained and tested on the TIMIT dataset. Results obtained from the proposed model show that it was able to outperform similar studies with DNN architectures on the TIMIT dataset without performing any feature extraction or transformation to the input, making it suitable for online and real-time applications. Compared with other initialization schemes, the use of wavelets, and gammatone filters significantly improved the results in every task. Out of three wavelet functions experimented with, Morlet wavelets produced the best performance, indicating that not all wavelet functions are suitable for this task. The proposed system achieved an MAE of 4.91 cm for females and 5.18 cm for males in height estimation, and an MAE of 6.07 years for females and 5.36 years for males in age estimation. In terms of gender detection, the model achieved accuracy of 99.98%. These results were further improved when calculating speaker-level predictions, indicating that progressive estimation from continues speech can provide a more accurate predictions. Through this study, several combinations of prediction tasks were explored to evaluate the effect of multi-task and single-task predictions. Taking gender in consideration, the results indicated no improvement over the findings when gender information was included.

This study aimed to shed the light on the ability to estimate various speaker characteristics using deep learning with raw waveform; however, it is important to note that the data used in this study was not diverse enough to capture the variabilities that affect the human voice. A more representative dataset is needed to be able to perform speaker profiling in real-world scenarios.

6 REFERENCES

[1] G. Assunção, P. Menezes, and F. Perdigão, "Speaker awareness for speech emotion recognition," *International Journal of Online and Biomedical Engineering (iJOE),* vol. 16, no. 4, pp. 15–22, 2020. https://doi.org/10.3991/ijoe.v16i04.11870

[2] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in *Ninth International Conference on Spoken Language Processing*, 2006: Interspeech, pp. 2118–2121. https://doi.org/10.21437/Interspeech.2006-195

[3] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Communication,* vol. 121, pp. 16–28, 2020, https://doi.org/10.1016/j.specom.2020.03.008

[4] J. Laver and P. Trudgill, "Phonetic and linguistic markers in speech," *Social Markers in Speech,* vol. 1, p. 32, 1979.

[5] A. H. Poorjam and M. H. Bahari, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2014: IEEE, pp. 7–12. https://doi.org/10.1109/ICCKE.2014.6993339

[6] 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: INTERSPEECH pp. 2277–2280. https://doi.org/10.21437/Interspeech.2007-618

[7] A. Badr and A. Abdul-Hassan, "A review on voice-based interface for human-robot interaction," *Iraqi Journal for Electrical and Electronic Engineering,* vol. 16, no. 2, pp. 1–12, 2020. https://doi.org/10.37917/ijeee.16.2.10

[8] S. Galgali, S. S. Priyanka, B. Shashank, and A. P. Patil, "Speaker profiling by extracting paralinguistic parameters using mel frequency cepstral coefficients," in *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2015: IEEE, pp. 486–489. https://doi.org/10.1109/ICATCCT.2015.7456933

[9] A. A. Badr and A. K. Abdul-Hassan, "Estimating age in short utterances based on multiclass classification approach," *Computers, Materials & Continua,* vol. 68, no. 2, pp. 1713–1729, 2021, https://doi.org/10.32604/cmc.2021.016732

[10] 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, 2010, https://doi.org/10.1007/s10772-010-9064-2

[11] A. Badr and A. Abdul-Hassan, "CatBoost machine learning based feature selection for age and gender recognition in short speech utterances," *International Journal of Intelligent Engineering and Systems,* vol. 14, no. 3, pp. 150–159, 2021, https://doi.org/10.22266/ijies2021.0630.14

[12] 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*, 2013: IEEE, pp. 7552–7556. https://doi.org/10.1109/ICASSP.2013.6639131

[13] 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, https://doi.org/10.1109/LSP.2013.2295397

[14] A. A. Mallouh, Z. Qawaqneh, and B. D. 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, no. 8, pp. 2581–2593, 2018, https://doi.org/10.1007/s00521-017-2848-4

[15] O. Büyük and M. L. Arslan, "Combination of long-term and short-term features for age identification from voice," *Advances in Electrical and Computer Engineering,* vol. 18, no. 2, pp. 101–108, 2018. https://doi.org/10.4316/AECE.2018.02013

[16] 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. 22524–22530, 2018, https://doi.org/10.1109/ACCESS.2018.2816163

[17] 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)*, 2019: IEEE, pp. 6580–6584. https://doi.org/10.1109/ICASSP.2019.8683397

[18] 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.

[19] 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.

[20] 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. https://doi.org/10.3390/s21144785

[21] D. Kwasny and D. Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors (Basel),* vol. 21, no. 14, p. 4785, 2021, https://doi.org/10.3390/s21144785

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015: IEEE, pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[23] R. Ardila *et al.,* "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670,* 2019.

[24] A. Tursunov, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors,* vol. 21, no. 17, p. 5892, 2021. https://doi.org/10.3390/s21175892

[25] M. Ravanelli and Y. Bengio, "Speech and speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1812.05920,* 2018. https://doi.org/10.1109/SLT.2018.8639585

[26] T. Parcollet, M. Morchid, and G. Linares, "E2E-SINCNET: Toward fully end-to-end speech recognition," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 7714–7718. https://doi.org/10.1109/ICASSP40776.2020.9053954

[27] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Systems with Applications,* vol. 136, pp. 252–263, 2019. https://doi.org/10.1016/j.eswa.2019.06.040

[28] E. Loweimi, P. Bell, and S. Renals, "On Learning Interpretable CNNs with Parametric Modulated Kernel-Based Filters," in *INTERSPEECH*, 2019, pp. 3480–3484. https://doi.org/10.21437/Interspeech.2019-1257

[29] T. Li *et al.*, "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 52, no. 4, pp. 2302–2312, 2021. https://doi.org/10.1109/TSMC.2020.3048950

[30] A. A. Alsalihi, H. K. Aljobouri, and E. A. K. ALTameemi, "GLCM and CNN deep learning model for improved MRI breast tumors detection," *International Journal of Online & Biomedical Engineering,* vol. 18, no. 12, 2022. https://doi.org/10.3991/ijoe.v18i12.31897

[31] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition,* vol. 90, pp. 119–133, 2019. https://doi.org/10.1016/j.patcog.2019.01.006

[32] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "An efficient auditory filter bank based on the gammatone function," *Annex C of the SVOS Final Report,* 1988.

[33] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica United with Acustica,* vol. 88, no. 3, pp. 433–442, 2002.

[34] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society,* vol. 79, no. 1, pp. 61–78, 1998. https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2

[35] M. Farge, "Wavelet transforms and their applications to turbulence," *Annual Review of Fluid Mechanics,* vol. 24, no. 1, pp. 395–458, 1992. https://doi.org/10.1146/annurev.fl.24.010192.002143

[36] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n,* vol. 93, p. 27403, 1993. https://doi.org/10.6028/NIST.IR.4930

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imageNet classification," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123

7 AUTHORS

Umniah Hameed Jaid is a doctoral candidate at the Computer Science Department in University of Technology and an assistant lecturer at the college of science, computer science department, at the University of Baghdad, Iraq. She finished her bachelor's in computer science in the department of computer science at the University of Baghdad and earned her master's degree in computer science from the College of science and Engineering, Department of Computer science, University of Glasgow, UK (email: umniah.h@sc.uobaghdad.edu.iq).

Alia Karim AbdulHassan is a professor and the dean of the Computer Science Department at the University of Technology, Baghdad, Iraq. She earned her bachelor's, master's, and doctoral degrees in computer science from the University of Technology. Her research interests include soft computing, green computing, AI and data mining (email: alia.k.abdulhassan@uotechnology.edu.iq).