Engineering and Technology Journal
Journal homepage: engtechjournal.org

Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks

Ameer A. Badr [a*], **Alia K. Abdul-Hassan** [b]

[a] Department of Computer Science, University of Technology, Baghdad, Iraq. College of Managerial and Financial Sciences, Imam Ja'afar Al-Sadiq University, Salahaddin, Iraq. cs.19.53@grad.uotechnology.edu.iq

[b] Department of Computer Science, University of Technology, Baghdad, Iraq.
110018@uotechnology.edu.iq

*Corresponding author.

Submitted: 29/10/2020 Accepted: 06/02/2021 Published: 25/03/2021

K E Y W O R D S

Gated-Recurrent Unit (GRU),
Linear Discriminant Analysis (LDA),
Speaker age estimation,
Statistical functional,
VoxCeleb1 dataset.

A B S T R A C T

Recently, age estimates from speech have received growing interest as they are important for many applications like custom call routing, targeted marketing, or user-profiling. In this work, an automatic system to estimate age in short speech utterances without depending on the text is proposed. From each utterance frame, four groups of features are extracted and then 10 statistical functionals are measured for each extracted dimension of the features, to be followed by dimensionality reduction using Linear Discriminant Analysis (LDA). Finally, bidirectional Gated-Recurrent Neural Networks (G-RNNs) are used to predict speaker age. Experiments are conducted on the VoxCeleb1 dataset to show the performance of the proposed system, which is the first attempt to do so for such a system. In gender-dependent system, the Mean Absolute Error (MAE) of the proposed system is 9.25 years, and 10.33 years, the Root Mean Square Error (RMSE) is 13.17 and 13.26, respectively, for female and male speakers. In gender_ independent system, the MAE of the proposed system is 10.96 years, and the RMSE is 15.47. The results show that the proposed system has a good performance on short-duration utterances, taking into consideration the high noise ratio in the VoxCeleb1 dataset.

How to cite this article: A. A. Badr, and A. K. Abdul-Hassan, "Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks," Engineering and Technology Journal, Vol. 39, Part B, No. 01, pp. 129-140, 2021.
DOI: https://doi.org/10.30684/etj.v39i1B.1905

1. INTRODUCTION

The speech contains valuable linguistic context information as well as paralinguistic information about speakers such as identity, emotional state, gender, and age [1]. Automatic recognition of this kind of information can guide Human-Computer Interaction (HCI) systems to adapt automatically to different user needs [2].

Automatic age estimation from speech signals has a variety of forensic and commercial applications. For example, in targeted internet advertising, information about the accent, language, gender, and age of the user can help to provide suitable services because of the significant increase of vocal interaction between user-company and user-computer over the past decades. Estimation of speaking age is also required in many forensic scenarios such as threatening calls, kidnapping, and falsified alarms to help identify criminals, e.g. shorten the number of suspects. Automatic age estimation may also be used for effective call diverting in call centers [1,3].

For several reasons, estimating a speaker's age is considered a difficult issue. First, it is difficult to estimate speaker's age using machine learning methods that work with discrete labels because the speaker's age is a continuous variable. Second, there is usually a difference between a speaker's age as perceived, namely the perceptual age, and their actual age, that is, the chronological age. Third, there are very few publicly available data sets labeled with age that have a sufficient number of speech utterances from a variety of age groups. Finally, Speech includes significant intra-age variability due to speech content, identity, speaking style, gender, emotional states, weight, height, etc., which makes the speakers of the same age sound different [1,4].

In this work, the analysis is performed on the VoxCeleb1 dataset, which is the first attempt to do so for such a system. There are several reasons to use this dataset in this work as it includes various background noise and varied utterance duration because it includes 100,000 YouTube utterances for 1,251 celebrities, the dataset is gender-balanced, and the speakers span a wide range of different ethnicities, accents, professions, and ages.

The primary contributions of the present work are highlighted and summarized as follows:

1) Combining four feature groups which are Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Subband Centroids (SSCs), Linear Predictive Coefficients (LPCs), and Formants to extract 150- dimensional feature vectors from each utterance.
2) Measuring 10 statistical functionals for each extracted feature dimension to achieve the greatest possible gain from each feature vector.
3) Using the Linear Discriminant Analysis (LDA) as a supervised dimensionality reduction method.
4) Suggesting a new age labeling to make the VoxCeleb1 dataset appropriate for the speaker's age estimation problem.
5) Exploring the impact of using bidirectional Gated-Recurrent Neural Networks (G-RNNs) in age prediction from short utterances.

The rest of this work is organized according to the following. The proposed system related works are presented in section two. Section three examines the theoretical background of G-RNN concepts in addition to feature types. Section four deals with the proposed methodology. The results of simulations and experiments are shown in section five. Finally, section six set out the work conclusions and future works.

2. RELATED WORKS

This work focuses mainly on speaker age estimation from short utterances, there are some previous works have concerned the study of such a system.

Spiegl [5] developed an acoustic feature set for the speaker's age estimation problem. As they stated, adding formant, pitch, and prosodic features to the MFCCs features leads to relative reductions of the Mean Absolute Error (MAE) between 4-20%. Their work focused less on comparing selected features but more on developing a feature vector that integrates many different cepstral, spectral and prosodic parameters to achieve a low error rate. Their work was evaluated on Florida Vocal Aging Database (UF-VAD). Their best result was 9.5 and 10.1 MAE for females and males respectively. Bahari [4] proposed a speaker's age estimation approach when Hidden Markov Model (HMM) weight supervectors were modeled for each speaker. Then, the dimensionality of input space has been reduced by using Weighted Supervised Non-Negative Matrix Factorization (WSNMF). Finally, speaker's age has been estimated by using the Least Squares Support Vector Regressor (LS-SVR). Their results were conducted on Dutch corpus with 7.9 MAE for both males and females. Bahari [3] proposed a speaker age estimation approach when an i-vector was modeled for each speaker utterance. Then, a session variability compensation was made by using a Within-Class Covariance Normalization (WCCN) technique. Finally, speaker's age has been estimated by

using the Least Squares Support Vector Regressor (LS-SVR). Their work was evaluated on NIST 2008 and 2010 SRE databases. Their best result in long speech utterances (45 seconds) was 6.47 and 6.99 MAE for females and males respectively. While for testing on short speech utterances (10 seconds), their best results were 8.5 and 8.27 MAE for females and males respectively. Grzybowska [6] propose an approach for the speaker age estimation problem by applying the fusion of acoustic features regression and, i-vectors. They achieve a significant enhancement of 12.6% comparing to the i-vector system. Their work was evaluated on the aGender database. Their best result was 9.77 and 10.63 MAE for females and males respectively. Zazo [7] proposed a novel speaker age estimation system based on Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN). Their system was able to handle short utterances. Their system was evaluated on NIST 2008 and 2010 SRE databases. Their best result in long speech utterances was 6.23 and 7.31 MAE for females and males respectively. While for testing on short speech utterances (10 seconds), their best results were 6.97 and 7.79 for females and males respectively. Table I demonstrates a summary of the related works mentioned above with the used methodology, datasets and achieved results.

TABLE I: The summary of the related works, their methodology, used databases and achieved results

Authors	Dataset	Methodology	Average Duration (sec)	MAE	
				Female	Male
Grzybowska [6]	aGender	i-vector, LS-SVR	3	9.77	10.63
Spiegl [5]	UF-VAD	MAX R, SVR	n/a	9.50	10.10
Bahari [3]	NIST SRE	i-vector, WCCN, LS-SVR	5	9.51	8.99
			10	8.5	8.27
Bahari [4]	Dutch corpus	HMM, WSNMF, LS-SVR	n/a	7.90	7.90
Zazo [7]	NIST SRE	i-vector, Neural Networks	5	9.85	10.82
			10	9.56	10.69
		LSTM-RNN	5	7.44	8.29
			10	6.97	7.79
The Proposed	VoxCeleb1	Statistical Functionals, LDA, G-RNN	8	9.25	10.33

Unlike all previous works that relied on HMM or i-vector for speaker modeling which are time-consuming, the proposed work depended on a simple feature fusion with exploiting the gain of statistical functionals as well as LDA to produce a simple informative feature vector (i.e. 30-dimensions). On the other hand, the proposed age estimation system train and tests on short utterances (i.e. 8 seconds) contrary to works in [3,7] which are train on long utterances and test on a short one. These two points demonstrate the ability of the proposed system to work with interaction or real-time applications.

3. FEATURE EXTRACTION METHODS

As mentioned before, the speech signal contains various types of paralinguistic information, e.g. speaker age. Features are determined at the first stage of all classification or regression systems, where a speech signal is transformed into measured values with distinguishing characteristics. Such methods used in this work are described below in brief.

I. Mel-Frequency Cepstral Coefficients (MFCCs)

Between all types of speech-based feature extraction domains, Cepstral domain features are the most successful ones, where a cepstrum is obtained by taking the inverse Fourier transform of the signal spectrum. MFCC is the most important method to extract speech-based features in this domain [8]. MFCCs greatness stems from the ability to exemplify the spectrum of

speech amplitude in a concise form. A speaker's voice is filtered by the articulator form of the vocal tract, such as the nasal cavity, teeth, and tongue. This shape affects the vibrational characteristics of the voice. If the shape is precisely controlled, this should give an accurate depiction of the phoneme being formed [9]. The MFCC features calculation steps are preemphasis, frame blocking and widowing, applying Fast Fourier Transform (FFT), applying mel-bank filter, taking the logarithm, and applying the discrete Cosine Transform (DCT). These steps are shown in Figure 1. At the end of these steps, one energy and 12 cepstral features are obtained [8,10].

II. Spectral Subband Centroids (SSCs)

SSC feature proposed by Paliwal [11] is intended to be a complement to the cepstral features in speech recognition. High sensitivity to additive noise distortion has been considered as one of the major problems with the cepstral-based features, the addition of white noise to the speech signals affects the spectrum of speech power at all frequencies, but in the higher amplitude (formant) portions of the spectrum, the effect is less noticeable. Therefore, to ensure the robustness of the feature, some formant-like features have to be investigated, SSC features are similar to the formant frequencies and can be easily and reliably extracted [11]. The entire frequency band (*0 to Fs/2*) is divided into N number of sub-bands for computation of SSCs, where *Fs* is the speech signal sampling frequency. SSCs are found by applying filter banks to the signal power spectrum and then calculating the first moment (centroid) of each subband. SSC of the m^{th} subband is calculated as seen in equation (1), where *Fs* is the sampling frequency, P(f) is the short-time power spectrum, ωm(f) is the frequency response of m^{th} bandpass filter and γ is the parameter controlling the dynamic range of the power spectrum [12].

$$C_m = \frac{\int_0^{Fs/2} f\ \omega m(f) P\gamma(f) df}{\int_0^{Fs/2} f\ \omega m(f) P\gamma(f) df} \quad (1)$$

III. Linear Predictive Coefficients (LPCs)

LPCs are techniques developed to analyze speech. The idea behind this is to model the production of speech as an additive model consisting of a source and a filter with one or more resonant frequencies. The source corresponds to the vocal folds' primary vibrations, and the filter is due to the vocal tracts' shapes and movements, that is, the throat, the tongue, and the lips [13]. By predicting a formant, LPC analyzing decided on a signal formant called inverse filtering, then estimated intensity and frequency from the residue speech signal. Because the speech signal has many time-dependent variations, the estimate will cut a signal called a frame. The procedure for obtaining the LPC coefficient is shown in Figure 2 [14].

IV. Formant based Features

The vocal tract shape contains much useful information, and its representation has been used widely in many speeches related applications. The formants, a representation of the vocal tract resonances, can be modeled with LPC [15]. Formants are nothing but the peaks of the spectral spectrum of the voice. In speech phonetics, formant frequencies are human vocal tract acoustic resonance which is measured as an amplitude peak in the frequency spectrum of the sound. In acoustics, the formants are referred to as the peak in the sound envelope and/or the resonance in the sound sources, as well as the sound chambers. Procedure to get formant features shown in Figure 3 [16].

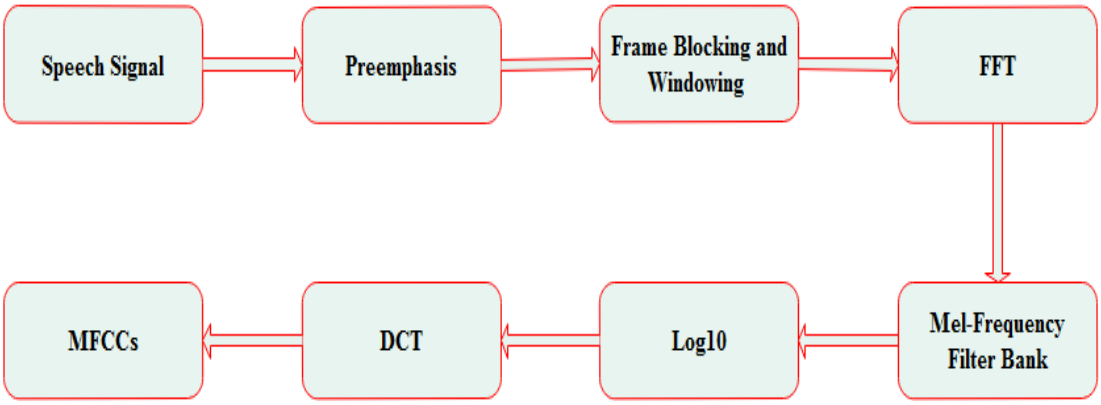

Figure 1: Mel frequency cepstral analysis [10].

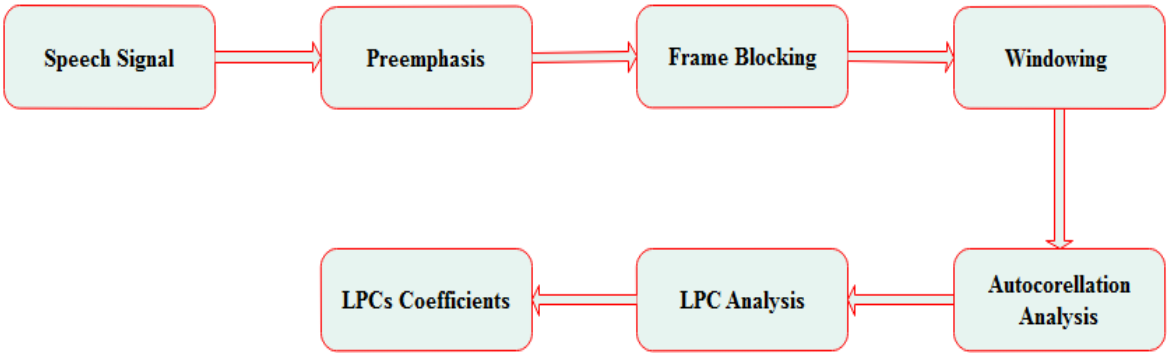

Figure 2: The LPC method [14].

Figure 3: Formant frequency detection process [16].

4. Dimensionality Reduction Method

Predominantly, there is some redundancy in the extracted high-dimensional features. Subspace learning can be used to eliminate these redundancies by further processing extracted features to reflect their semantic information better. Among all dimensionality reduction techniques, LDA is a very common supervised technique for dimensionality reduction problems as a preprocessing step for machine learning and pattern classification applications. The LDA technique aims at projecting the original data matrix in a lower-dimensional space. Three steps were needed to achieve this aim. The first step is to calculate the between-class variance (i.e. the distance between the means of different classes. The second step is to calculate the within-class variance, which is the distance between the mean and the samples of each class. The third step is to construct the lower-dimensional space which minimizes the within-class variance and maximizes the between-class variance [17].

5. Bidirectional Recurrent Neural Networks (Gated Recurrent Unit)

The deep learning is used for solution and process the complex problems recently [18]. RNNs are deep neural networks (DNNs), in which connections can form a cycle of direction. RNNs can be a good approach to modeling time sequences because it exhibits a dynamic temporal behavior [7]. At the current time step, RNN can map the output from the entire memory of previous inputs, which is significant for the processing of speech signals with a close timing relationship. Standard RNN can access information only from past inputs. By processing data in both directions, Bidirectional RNN can however access past and future contexts [19].

As seen in Figure 4, the main idea of the bidirectional RNN is to split the standard RNN hidden layer into the forward states part and the backward state's part. There is no connection between the

two parts, but both parts connect to the same input layer and the same output layer. The distinction between the two parts is that forward states are determined by past inputs along a positive time axis, and backward states are measured by future inputs along the reverse time axis [19].

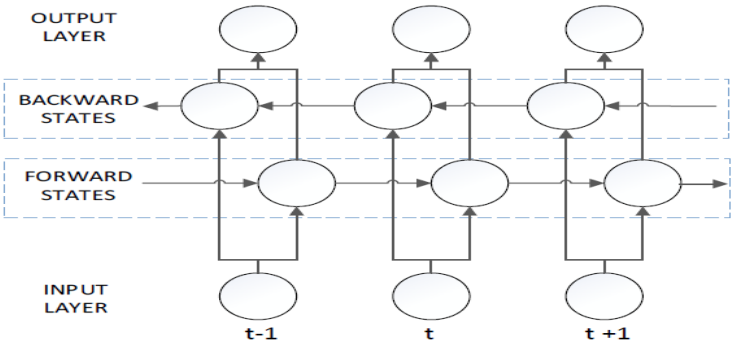

Figure 4: The structure of bidirectional RNN [19].

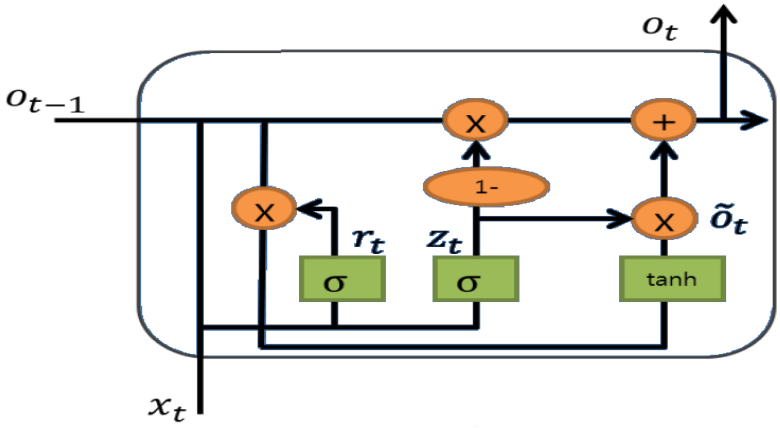

Figure 5: The structure of the GRU cell unit [20].

A study of RNN error flow shows that the standard RNN structure can only maintain short-term memory because of the vanishing gradient problem [7]: "the gradient of the total output error concerning previous inputs quickly vanishes as the time lags between relevant inputs, making them difficult to train and apply to real-life applications".

Gated Recurrent Unit (GRU) was proposed in [21] to adapt recurrent blocks to capture dependence on different time scales [19]. In the GRU cell unit, one gate controller controls both forget and input gates. When Z_t is 0, the forget gate is closed and the input gate is opened, whereas Z_t outputs 1, the forget gate is opened and the input gate is closed. In this way, the time step input is deleted whenever the previous (*t - 1*) memory is stored. The reset gate determines the way the new input is combined with the previous memory, and the update gate decides how much of the previous memory information is retained to calculate the new state as seen in Figure 5, and the following equations [20]:

$$r_t = \sigma(W_{xr}^T . X_t + W_{or}^T . o_{t-1} + b_r) \quad (2)$$

$$z_t = \sigma(W_{xz}^T . X_t + W_{oz}^T . o_{t-1} + b_z) \quad (3)$$

$$\tilde{o}_t = \tanh(W_{x\tilde{o}}^T . X_t + W_{o\tilde{o}}^T . (r_t \otimes o_{t-1}) + b_{\tilde{o}}) \quad (4)$$

$$o_t = z_t \otimes o_{t-1} + (1 - z_t) \oplus \tilde{o}_t \quad (5)$$

Where σ is sigmoid activation function, W_{xr}, W_{xz}, $W_{x\tilde{o}}$ denote the corresponding connected input vector weight matrices, W_{or}, W_{oz}, $W_{o\tilde{o}}$ represent the previous time step weight matrices, and b_r , b_z, $b_{\tilde{o}}$ are bias. $\otimes$ is a tensor product, $\oplus$ is a circular plus.

6. THE PROPOSED SPEAKER AGE ESTIMATION SYSTEM

In speaker's age estimation, one can give a training dataset of speech recordings $S^{train} = \{(X_1, Y_1), \dots, (X_s, Y_s)\}$. In this set, the S^{th} utterance of the training dataset and its corresponding speaker age denote X_s and Y_s respectively. The target is to design an estimator function, such that for an unseen utterance X^{test}, the actual age of the speaker is accurately predicted. As in Figure 6, the methodology of this work consists of five main stages which are feature extraction, statistical functionals measured, features normalization, dimensionality reduction, and bidirectional G-RNN regression.

Where at the beginning, appropriate features will be extracted from each speaker's utterance, to be followed by features scaling to fall within a smaller range by using normalization techniques. Then, by using the dimensionality reduction method, the high dimensional features will be transformed into more meaningful low dimensional features. Finally, a bidirectional G-RNN regressor is used to predict the speaker actual age.

I. Utterance based Features Extraction

As mentioned earlier, the issue of estimating a speaker's age is a difficult one where the extracted features need to be speaker-independent. For that, four groups of features have been fusion in this work in which the estimation errors from these different feature groups are complementary, which allows the combination of estimates from these feature groups to further improve the system performance. In the beginning, each speaker's utterance is split into frames with a window size of 250 milliseconds and a frameshift of 10 milliseconds to ensure that each frame contains robust information. Then, four groups of features have been extracted from each utterance frame, which are MFCC (i.e. 20-dimensions with its first and second derivative), LPC (i.e. 20-dimensions with its first and second derivative), SSC (i.e. 26-dimension), and formants (i.e. F1, F2, F3, and F4.). The total dimensions of the extracted features in this stage are 150 as seen in Figure 7.

Figure 6: The proposed speaker age estimation system.

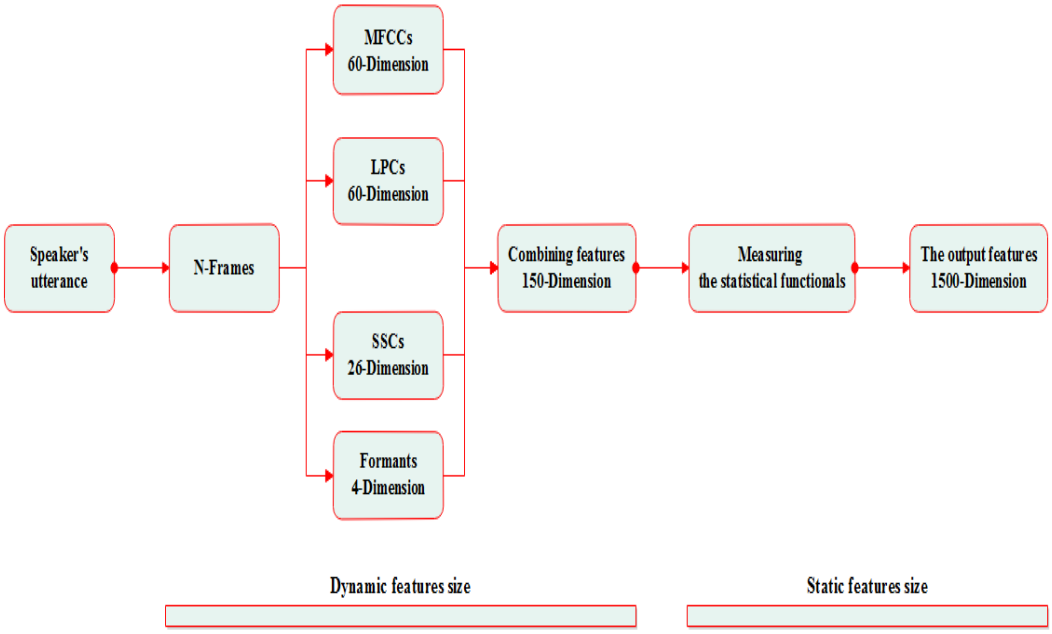

Figure 7: The proposed features fusion.

II. Statistical Functionals Measured

To override the issue of varying features size between different speaker utterances as well as to achieve the greatest possible gain from each feature dimension, the features with dynamic size extracted from the previous stage (i.e. 150-dimension) have been turned into features with static size by measuring 10 statistical functionals for each dimension. These statistical functionals include mean, min, max, median, stander deviation, skewness, kurtosis, first quantile, third quantile, and interquartile range (Iqr). The total output of features dimension in this stage is 1500 as seen in Figure 7.

III. Features Normalization using Z-Score

The expression of features in smaller units will result in a wider range for these features and will, therefore, tend to give these features greater effect. Normalization involves transforming the data to fall within a smaller or common range. Therefore, due to the great usefulness of the normalization process in machine learning methods, the 1500-dimensional features, extracted from the previous stage, will be normalized by using the z-score method [22].

IV. Dimensionality Reduction using LDA

At this stage, LDA takes as its input a set of 1500-dimensional normalized features grouped into labels. Then, an optimal transformation is found which maps these input features into a lower-dimensional space while preserving the label structure. It maximizes the between-label distance and at the same time minimizes the within-label distance, thus achieving maximum discrimination. The output of this stage is chosen empirically to be a 30-dimensional feature vector that contain the most important information to estimate speaker's age accurately from their voice.

V. Bidirectional G-RNN Regressor

As in Figure 8, the proposed regressor consists of two bidirectional GRU hidden layers followed by a flatten layer and finally a dense output layer. For each bidirectional GRU hidden layer, the bidirectional merge mode used is average, the activating used is tanh, the recurrent activation used is sigmoid, and unit size is 256, 128 respectively. The flatten layer is used to flatten the output of the bidirectional GRU hidden layer. A single dense with a linear activation function has been chosen to be the output layer that performs the regression to the target age.

To train the proposed deep network, Mean Squared Error (MSE) [23] has been used as a loss function, Adam algorithm [24] with a learning rate of 0.0001 has been used as an optimizer, and the number of epochs have been chosen depending on early stooping criteria. The topology of the proposed deep network is selected empirically and also by taking the topology in [7] into account.

7. VOXCELEB1 DATASET DESCRIPTION

The VoxCeleb1 dataset has been collected through automatic pipelines from open-source media, it is a multilingual dataset contains 1,250 speakers in text-independent scenarios. All utterances are encoded with 16-bit resolution at a 16 kHz sampling rate. The dataset includes diverse background noise and varied utterance duration because it comprises various utterances of celebrities from YouTube [25]. This dataset has been used in speaker identification as in [26] and speaker verification as in [27].

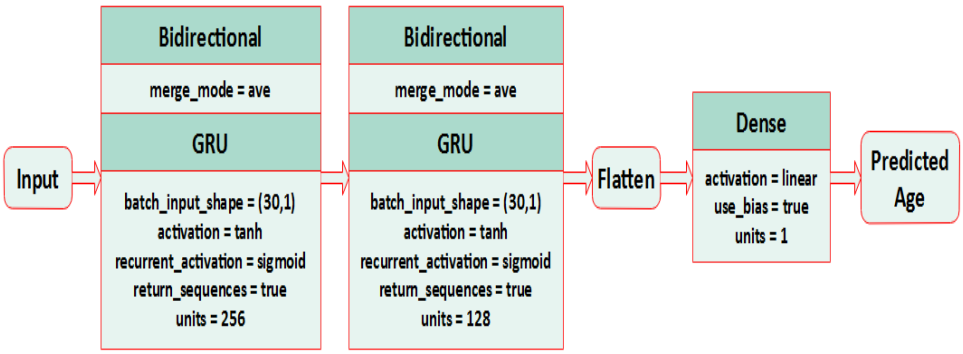

Figure 8: The proposed bidirectional G-RNN topology.

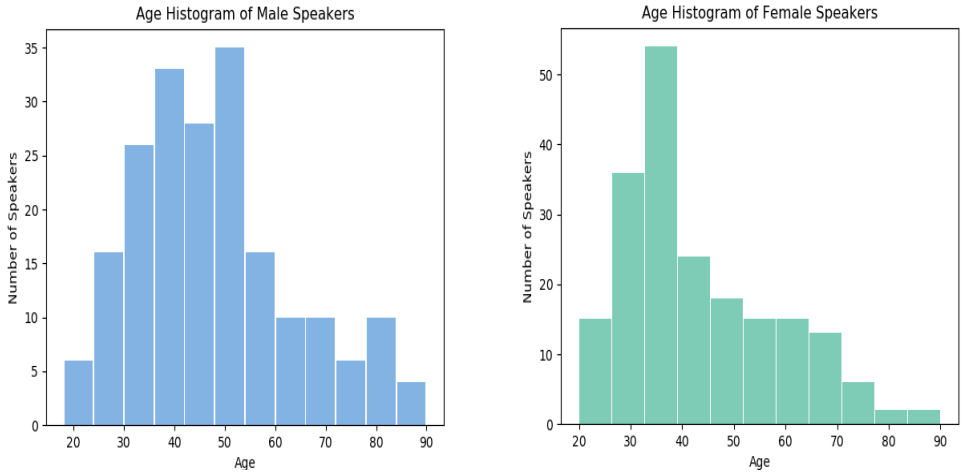

Figure 9: Age histogram of selected speakers.

In this work, a new age labeling has been proposed to make the VoxCeleb1 dataset appropriate for the speaker's age estimation problem, which is the first attempt to use it for such a system. 400 speakers have been chosen with 25 utterances for each resulting in a total of 10,000 utterances to get sufficient data to train and test the system. Half of the speakers are males, and the other half are females. The average utterance duration is 8.2, 8.4 seconds for males and females respectively. The statistics of the dataset is given in Table II; the age histogram of the dataset speakers is shown in Figure 9.

TABLE II: The VoxCeleb1 dataset statistics.

Gender	No. Speakers	Age (year)			
		Minimum	Maximum	Mean	Standard deviation
Males	200	18	90	47.58	15.73
Female	200	20	90	43.28	14.57
Male + Females	400	18	90	45.43	15.15

8. EXPERIMENTAL RESULTS AND DISCUSSIONS

The dataset used in this work was described in detail in this section, and also the results of the experiments were explained and discussed. The performance of the proposed age estimation system is assessed by two objective measures used in previous works.
MAE is calculated according to Eq. (6) [7]; lower MAE means better performance. Root Mean Square Error (RMSE) is computed as in Eq. (7) [28]; lower RMSE means better performance.

$$MAE = \frac{1}{N} \sum_{n=1}^{N} \left| \check{Y}_n - Y_n \right| \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (Y_n - \check{Y}_n)^2}{N}} \quad (7)$$

Where N is the number of utterances, Y_n the ground truth for the age, $\check{Y}_n$ is the predicted age.

In order to train, test, and compare the proposed system consistently, the VoxCeleb1 dataset has been divided into two parts; the training set contains 4500 utterances for each male and female speakers, while the test set contains 500 utterances for each male and female speakers. To prevent overfitting during the training process, the overlapping of utterances in partitioning have been avoided. The proposed speaker age estimation system performance has been evaluated in terms of MAE and RMSE in order to show the effectiveness of the proposed system on short speech utterances. The experiment has been conducted in gender-dependent system as well as gender-independent system. Table III shows the results of this experiment.

TABLE III: MAE (years) and RMSE of the proposed system on short utterances from the VoxCeleb1 dataset.

The Proposed System	Gender	MAE	RMSE
Gender-Dependent	Male	10.33	13.26
	Female	9.25	13.17
Gender-Independent	Male + Female	10.96	15.47

As seen in Table III, the effectiveness of the proposed age estimation system has been evaluated using two measures: MAE and RMSE. In the proposed gender-dependent system, the MAE and RMSE of female speakers is slightly different from the MAE and RMSE of male speakers and that because of using a balanced number of utterances. On the other hand, the MAE and RMSE is slightly increased in gender-independent system. This shows the high efficiency of the proposed system in gender-independent considerations.

Bahari [3] stated that age estimation problems require a long duration of utterance to be effective (e.g. 45 seconds). The achieved MAE (i.e. years) by the proposed system is somewhat bigger when comparing to the related works on short utterances. This is caused by wider age range for the VoxCeleb1 dataset as seen in Figure 9 and that lead to greater MAE.

9. CONCLUSION AND FUTURE WORKS

An automatic system to estimate age in short speech utterances without depending on the text has been proposed in this work. Four groups of features are extracted and then 10 statistical functionals are measured for each extracted dimension of the features to produce a 1500-dimensional feature vector, which then reduced into a 30-dimensional informative feature vector by using LDA. Bidirectional G-RNNs are used to predict speaker's age accurately. Experimental results have clearly shown the effectiveness of the proposed system using the VoxCeleb1 dataset either in gender-dependent system with MAE of 9.25 and 10.33 for female and male speakers respectively or in gender-independent system with MAE of 10.96. For future works, one can use another type of dimensionality reduction method such as an auto-encoder neural network, and using an ensemble regressor on top of the deep RNNs to predict age with fewer errors.

References

[1] P. G. Shivakumar, M. Li, V. Dhandhania, and S. S. Narayanan, "Simplified and supervised i-vector modeling for speaker age regression," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., no. May, pp. 4833–4837, 2014, doi: 10.1109/ICASSP.2014.6854520.

[2] M. Li, C. S. Jung, and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010, no. January, pp. 2826–2829, 2010.

[3] M. H. Bahari, M. McLaren, H. Van Hamme, and D. A. Van Leeuwen, "Speaker age estimation using i-vectors," Eng. Appl. Artif. Intell., vol. 34, no. January 2018, pp. 99–108, 2014, doi: 10.1016/j.engappai.2014.05.003.

[4] M. H. Bahari and H. Van Hamme, "Speaker age estimation using Hidden Markov Model weight supervectors," 2012 11th Int. Conf. Inf. Sci. Signal Process. their Appl. ISSPA 2012, no. July, pp. 517–521, 2012, doi: 10.1109/ISSPA.2012.6310606.

[5] W. Spiegl et al., "Analyzing features for automatic age estimation on cross-sectional data," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. May 2014, pp. 2923–2926, 2009.

[6] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 08-12-Sept, no. September, pp. 1402–1406, 2016, doi: 10.21437/Interspeech.2016-1118.

[7] 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, doi: 10.1109/ACCESS.2018.2816163.

[8] A. A. Badr and A. K. Abdul-Hassan, "A Review on Voice-based Interface for Human-Robot Interaction," Iraqi Journal for Electrical And Electronic Engineering, vol. 16, no. 2, pp. 91–102, 2020.

[9] R. A. Mohammed, N. F. Hassan, and A. E. Ali, "Arabic Speaker Identification System Using Multi Features," Eng. Technol. J., vol. 38, no. 5A, pp. 769–778, 2020, doi: 10.30684/etj.v38i5a.408.

[10] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," Appl. Acoust., vol. 98, pp. 52–61, 2015.

[11] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), 1998, vol. 2, pp. 617–620.

[12] S. V Chougule and M. S. Chavan, "Speaker Recognition in Mismatch Conditions: A Feature Level Approach," Int. J. Image, Graph. Signal Process., vol. 9, no. 4, pp. 37–43, 2017, doi: 10.5815/ijigsp.2017.04.05.

[13] J. Sueur and others, Sound analysis and synthesis with R. Springer, 2018.

[14] W. S. Mada Sanjaya, D. Anggraeni, and I. P. Santika, "Speech Recognition using Linear Predictive Coding (LPC) and Adaptive Neuro-Fuzzy (ANFIS) to Control 5 DoF Arm Robot," J. Phys. Conf. Ser., vol. 1090, no. 1, 2018, doi: 10.1088/1742-6596/1090/1/012046.

[15] D. Hutchison and J. C. Mitchell, Affective Computing and Intelligent Interaction, vol. 137 AISC. 2012.

[16] [A. A. Khulage and B. V Pathak, "Analysis of speech under stress using Linear techniques and Non-Linear techniques for emotion recognition system," ArXiv, vol. abs/1207.5, 2012.

[17] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," Ai Commun., vol. 30, pp. 169–190, May 2017, doi: 10.3233/AIC-170729.

[18] A. A. Abdulhussein and F. A. Raheem, "Hand Gesture Recognition of Static Letters American Sign Language (ASL) Using Deep Learning," Eng. Technol. J., vol. 38, no. 6A, pp. 926–937, 2020, doi: 10.30684/etj.v38i6a.533.

[19] Y. Tang, Y. Huang, Z. Wu, H. Meng, M. Xu, and L. Cai, "Question detection from acoustic features using recurrent neural network with gated recurrent unit," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2016-May, pp. 6125–6129, 2016, doi: 10.1109/ICASSP.2016.7472854.

[20] H. M. Lynn, S. B. Pan, and P. Kim, "A Deep Bidirectional GRU Network Model for Biometric Electrocardiogram Classification Based on Recurrent Neural Networks," IEEE Access, vol. 7, pp. 145395–145405, 2019, doi: 10.1109/ACCESS.2019.2939947.

[21] K. Cho, B. V Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," in SSST@EMNLP, 2014.

[22] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.

[23] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures," IEEE Signal Process. Mag., vol. 26, no. 1, pp. 98–117, 2009, doi: 10.1109/MSP.2008.930649.

[24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," CoRR, vol. abs/1412.6, 2015.

[25] A. Nagraniy, J. S. Chungy, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2017-Augus, pp. 2616–2620, 2017, doi: 10.21437/Interspeech.2017-950.

[26] A. Hajavi and A. Etemad, "A deep neural network for short-segment speaker recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 2878–2882, 2019, doi: 10.21437/Interspeech.2019-2240.

[27] J. W. Jung, H. S. Heo, J. H. Kim, H. J. Shim, and H. J. Yu, "RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 1268–1272, 2019, doi: 10.21437/Interspeech.2019-1982.

[28] 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, pp. 6580–6584, doi: 10.1109/ICASSP.2019.8683397.