IMPROVING ROBUSTNESS OF AGE AND GENDER PREDICTION BASED ON CUSTOM SPEECH DATA

Veera Vignesh Kandasamy and Anup Bera

Accenture Solutions India Pvt Ltd, India

ABSTRACT

With the increased use of human-machine interaction via voice enabled smart devices over the years, there are growing demands for better accuracy of the speech analytics systems. Several studies show that speech analytics system exhibits bias towards speaker demographics, such age, gender, race, accent etc. To avoid such a bias, speaker demographic information can be used to prepare training dataset for the speech analytics model. Also, speaker demographic information can be used for targeted advertisement, recommendation, and forensic science. In this research we will demonstrate some algorithms for age and gender prediction from speech data with our custom dataset that covers speakers from around the world with varying accents. In order to extract speaker age and gender from speech data, we've also included a method for determining the appropriate length of audio file to be ingested into the system, which will reduce computational time. This study also identifies the most effective padding and cropping mechanism for obtaining the best results from the input audio file. We investigated the impact of various parameters on the performance and end-to-end implementation of a real-time speaker age and gender information extraction system. Our best model has a RMSE value of 4.1 for age prediction and 99.5% for gender prediction on custom test dataset.

KEYWORDS

Age and Gender prediction, Data Bias, Speech Analytics, CNN, LSTM, Wav2Vec.

1. INTRODUCTION

Speech analytics is the most predominant subject that has been used in various industries as a tool to understand consumers, predict human behaviour and create content. The speech analytics has various use cases like contact centre operations (Litvinov et.al)[1] customer service improvement (Scheidt et.al) [2], new language learning and mental health screening (Yiling Li) [3]. In all of these use cases, speech to text, speech sentiment, text to speech technologies are considered primarily. Speech analytics is the practice of collecting and analysing voice data with the use of speech technologies to improve the user experience. Speech analytics comprises of various functions, but it is subjected to bias as all other machine learning algorithms. These biases might occur at any stage in the speech analysis process. In case of call centre applications, the speech to text models used, might be trained on some data irrespective of speaker demographics taken into consideration. When speech to text is performed on audio calls for a specific age group or gender for which training data is not readily available, this can lead to bias issues. In this study, we are attempting to understand speaker characteristics in terms of age and gender for a certain set of audio samples that represent the overall population. If the initial speech-to-text model was trained using data with skewed sampling for age group and gender, it will be difficult for the system to

David C. Wyld et al. (Eds): SIGEM, MLTEC, SEAPP, ITCON, NATL, FUZZY, CSEA - 2022
pp. 69-83, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.122005

maintain accuracy across all age groups and gender. To avoid bias of a model, it is important to prepare a training dataset which has equal representation of all age groups and gender. To achieve this, there should be some frameworks for extraction of the age and gender information from a speech data. In this study we will explore some of the deep learning approaches to predict the speakers age, age group and gender.

1.1. Unique Contribution of this Work are:

•In this study, we have combined multiple datasets like TIMIT, VCTK, NISP, and GMU that provide more generalization in terms of age distribution and gender of the speaker. This helps in making the solution more robust and generalised during the usage in real world problems

•As part of the research, the appropriate duration of the audio recording required to accurately assess the age and gender is determined. On the combined custom dataset, the 5s appears to provide better generalisation values based on performance and computational cost.

•In this study, the impact of padding and cropping input audio files on accuracy metrics is also investigated.

•Accurately predict age [discrete and continuous] and gender from the audio samples

- Weighing mechanisms used to better understand classes with lower representation (Classification Problem)

Our paper is organized as follows: Section 2 reviews the available literature on the selected topic and gives some theoretical background, section 3 contains the Research methodology as well as research questions section 4 contains the results based on our research, and section 5 concludes our research paper with future work.

2. THEORETICAL BACKGROUND AND LITERATURE REVIEW

This section briefly discusses some issues, and showcases literature related to devising the age and gender detection problem.

2.1. Use of Machine Learning Models to Detect Age and Gender

In Safavi et.al. [4] propose a hybrid model that makes predictions and classifies age, gender, and accent using different models. XGBoost, a decision tree based on ensemble technique, is used to classify gender, KNN for Age, and Random Forest for detecting the accent of a speaker. In Jasuja et.al. [5], implemented a gender specific user classification using a deep learning model based on the Multilayer Perceptron (MLP). The proposed model was trained with various parameters and produced an MLP model with 96% accuracy in the test dataset. These predictions were showcased on a very clean dataset which would not be in case of real-world situations like a call centre.

2.2. Use of Deep Learning Models to Detect Age and Gender

Buyukyilmaz [6] proposed a multi-layer perceptron deep learning model helped detect gender based on the auditory characteristics of sound and voice. The classification model was 96.74 percent accurate. Maka and Dziurzanski [7] used a dataset containing 438 men and 192 women (indoor and outdoor listening scenes) and used 630 speakers in experiments on gender

identification issues in different language contexts. They found that nonlinear smoothing improved classification accuracy by 2% to become 99.4% in total. Tursunov [8] propose a novel CNN model with Multi-Attention Module (MAM) for age and gender classification using speech spectrograms from the input speech signals. Sánchez-Hevia [9] identifies the available types of DNN techniques to simultaneously detect age and gender in the context of Interactive Voice Response (IVR) systems. The results reveal that the larger the network, the better the results, but the improvement is insufficient to justify the additional processing cost.

3. RESEARCH METHODOLOGY

3.1. Dataset Description

In order to make the Age and Gender prediction model robust we have collected various available datasets. The following is a brief description about each of them

NISP Dataset [10]**:** The NISP dataset consists of audio samples from 345 speakers (60% male and 40% female) each contributing about 4-5 minutes of data. For our use case we are considering only English sentences spoken by the speakers. The dataset contains the exact age of the speaker, and the distribution ranges from teens to forties.

TIMIT Dataset [11]**:** The dataset includes recordings of over 600 American speakers from eight different dialects with 70% male and 30% female. The dataset includes the speaker's exact age, which ranges between the twenties and the seventies.

VCTK Corpus [12]**:** The VCTK dataset is made up of audio recordings of 110 different speakers reading 400 different phrases from newspapers. It contains the exact age of the speaker, ranging from teens to thirties, with male speakers with 57% of the data and female speakers of 43%. The audio was recorded in a professional noise free setting with two distinct microphones at 24 bits and then down sampled to 48KHz stored as FLAC. Only the audio from microphone 2 is being considered.

GMU Corpus [13]**:** The dataset contains speakers from various part of the world with different accent read the same English sentence. The latest version of the dataset contains around 2982 samples of such recordings from which we have removed synthesized audio.

Combined Dataset: In the individual datasets listed above, there is a significant imbalance in the data's male-female distribution, and each data set contains speakers only from a single region. To mitigate this, we must first create a bias-free dataset that covers the entire spectrum of speakers.

Table 1 lists the number of records from each of the dataset across different age groups. Since the datasets are in different formats, we built a pipeline to standardise the audio input that must be passed into our model. The choice of the parameters chosen to be standardized are discussed below.

Table 1. Dataset details across age groups

Dataset	**Infants**	**Teens**	**Twenties**	**Thirties**	**Forties**	**Fifties**	**Sixties**	**Seventies**	**Eighties**	**Nineties**	***Total***
GMU	2	320	1344	542	325	260	103	43	20	3	***2962***

NISP	0	688	11226	2222	556	0	0	0	0	0	***14692***
TIMIT	0	0	3630	1860	570	200	30	10	0	0	***6300***
VCTK	0	5456	37183	1234	0	0	0	0	0	0	***43873***
TOTAL	***2***	***6464***	***53383***	***5858***	***1451***	***460***	***133***	***53***	***20***	***3***	***67827***

3.2. Research Questions

Majority of the work in this study focuses on creating a balanced dataset towards sensitive variables, such as age and gender and to optimize the computational aspect with better prediction accuracy.

Following are the key questions answered in this research:

•How can we create a dataset that represents the entire range of speakers without bias?
•What is the effectiveness of the proposed data ingestion module over state of art age and gender detection models?
•What is the required audio length for understanding characteristics of age and gender accurately?
•How can we accurately predict age and gender using a single model?
•What is the impact of padding and cropping on the training and validation metrics?

3.3. Data Processing Framework

3.3.1. Data Standardization

We know from the last section that the data comes from different sources with different formats and sample rates different naming for the speaker's profile information. A Pipeline was created to convert all the datasets to wav format with 16KHz sampling rate in order to provide a balance between the training time and the accuracy of the model.

Table 2 show cases the list of datasets with their existing file type and other metadata.

All the datasets are converted into the standard format mentioned below.

•File Type: WAV
•Sampling Rate: 16000 KHZ
•Number of Channels: 1
•Bits Per sample: 16
•Encoding: PCM S

Table 2. Dataset Metadata description

Dataset Name	File Type	Sample Rate	Channels	Bits Per Sample	Encoding
TIMIT	WAV	16000	1	16	PCM_S
NISP	WAV	48000	1	16	PCM_S
GMU	WAV	44100	1	16	PCM_S
VCTK	FLAC	48000	1	16	PCM_S

3.3.2. Description of Dataset

From Table 1 we understand that the number of records in the infants and nineties are extremely low. For our use case, we are concentrating on speakers whose age range is in teens to fifties. We

also down sampled the number of records in the teens and twenties age range to avoid bias in the model towards these majority classes.

Table 3 showcases the final Combined Dataset prepared that includes the train and validation set from the various standard datasets. Both age group (Teens – Fifties) distribution and gender (Male & Female) distribution are shown in the same. While preparing the dataset, we made sure to include the best possible distribution that represents both men and women which is shown in Figure 1

Table 3. Age and Gender Dataset

Dataset	Teens	Twenties	Thirties	Forties	Fifties	Male	Female
GMU	274	169	469	262	224	685	713
NISP	582	1314	1889	487	0	2545	1727
TIMIT	0	301	1165	402	92	1494	466
VCTK	4683	4094	1050	0	0	3845	5982
Total	**5539**	**5878**	**4573**	**1151**	**316**	**8569**	**8888**

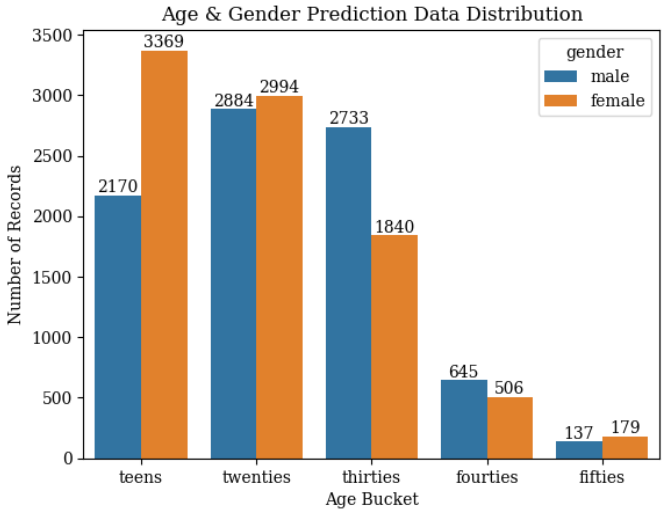

Figure 1. Age & Gender Dataset Distribution

3.3.3. Audio Duration processing

Audio can be thought of as a single-dimensional array of numbers. In order to process the input sequences, our model requires them to be the same length. This can be accomplished through the use of pre-processing techniques. A cut off duration can be defined based on which either padding or cropping can be done to the input audio. When our audio is shorter than the cut-off, we must pad the sequence data to the left or right to make it in desired length. Similarly, if the length of the audio exceeds the cut off, we crop it to the desired length on the right, left, or at random. The current literature places little emphasis on the ideal length of audio required for accurate age and gender prediction, owing to the dataset's limited distribution of audio length. In our study, we experimented with various audio lengths that would be required to optimise prediction accuracy as we have included data from multiple datasets. We also investigate the

impact of various padding and cropping strategies on the input audio. Further experiments were conducted following the selection of the ideal audio duration required.

3.4. Experimental Design

The following experiments were developed using the combined dataset created in order to find answers to the research's key questions.

1. Finding the ideal duration that would predict the age and gender of the speaker. Using the combined dataset we chose a model and applied by modifying the duration of the audio.
2. Finding the effective padding and cropping strategy for the selected audio duration. Padding and Cropping can be done either on the left or right or at random position of the audio.
3. Find the optimal the number of MFCC feature.
4. CNN model with MFCC & Wav2vec as the input features
5. LSTM model with MFCC & Wav2vec as the input features
6. CNN+LSTM model with MFCC & Wav2vec as the input features
7. Multi CNN+LSTM model with MFCC & Wav2vec as the input features.

Every experiment builds on the one before it, answering important questions at each stage and using the optimal parameters afterward. For instance: The ideal audio length for predicting age and gender is determined in experiment 1 and used to determine the padding and cropping method in experiment 2.

3.5. Feature Extraction

Feature extraction techniques are used to extract pertinent information from an audio source. Existing literature uses MFCC [9], Mel Spectrogram [14], Mel-filter banks [15] as input features, along with Pitch, Chroma, and Tonnetz as additional features for performing age and gender classification. It is established that MFCC provides resiliency to Noise in the dataset.

In the case of pretrained models like wav2vec [16] which is trained in an unsupervised manner on a different set of audio dataset can be fine-tuned based on the problem using transfer learning techniques. In these pretrained models feature extractions are accomplished internally; nonetheless, a certain format and sample rate are required to use these models. Internally wav2vec uses CNN encoder over each segment of the raw audio and learns by predicting the masked speech unit.

As part of our research, we have experimented with MFCC and Wav2Vec as the feature extraction technique.

3.6. Model Description

Our model takes in audio from the custom dataset as batches. The general flow as shown in Figure 2. consists of 3 steps. The feature extraction layer takes in two values either wav2vec or MFCC. In case of MFCC the best number is chosen from the result of previous experiment. The model layer consists of CNN, LSTM, CNN+LSTM, Multi CNN + LSTM. The experiments were carried out with various combination of feature and model type. The model is then able to predict three different values age, age group and gender of the speaker.

We have developed these models in python 3.9 [17] using Pytorch Lightning [18] which is a wrapper around Pytorch [19].

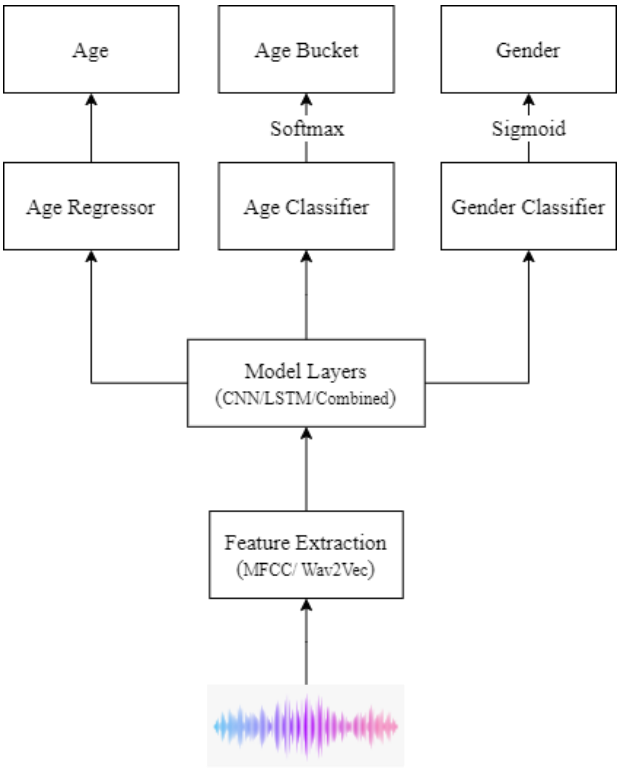

Figure 2. General Model Description

3.6.1. CNN Model

We have two kinds of CNN model to accommodate for variation in feature extractors. For Wav2Vec feature we pass in the raw audio and for MFCC based model we extract the MFCC and then pass it to the model. The Model layer consists of Conv1d Layer with ReLU and Batch Normalization. Soft Attention is applied at the end to reduce the dimension.

Output of the attention layer is then passed to the multi head setup where it predicts age, age group and gender of the speaker from the input. As age is regression problem, we use a single neuron as the last layer. The neurons in the final layer for age group correspond to the five age groups we have in our situation. We have taken a softmax of these values so as to get the probability of each class. While predicting gender as it's a binary classification problem we have one neuron as the last layer with sigmoid applied.

3.6.2. LSTM Model

LSTM models are known to work good for sequence prediction problems. Similar to CNN Model we have two variants of the model based on the feature extractor. Model layer consists of LSTM with soft attention. Output of the attention layer is then passed to the multihead classifier and regressor.

3.6.3. CNN+LSTM Model

We combined the aforementioned two models to increase complexity and, as a result, performance while enabling our model to pick up additional features. Model Layer consists of

Conv1d layer followed by LSTM and Soft Attention. Output of the soft attention is then used to predict the age, age group and gender.

3.6.4. Multi CNN+LSTM Model

In this Model we have experimented with varying Convolution stack(n*Conv1d+ReLU+BatchNorm) and LSTM with self-attention layers. The model is made more complex so that it can pick up new features and improve its performance in classification and regression tasks.

3.7. Evaluation of Performance

Our models predict the following outcomes

1. The speaker's exact Age - Regression Problem
2. The speaker's age group - Classification Problem
3. The speaker's Gender prediction - Classification Problem (Binary)

In the case of Age as a regression problem, our model should predict the continuous value, which should be compared to the ground truth label, with the goal of minimising the error.

In the case of a classification problem, gender has two values, male and female, whereas the age group has five values that the model must predict (teens, twenties, thirties, forties, and fifties). The classification problem's goal is to compare the predicted class to the ground truth label and make it predict the exact class as accurately as possible.

Regression Problem: In case of a regression problem, the RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) statistics are used, with lower values indicating better performance. Mathematically

$$RMSE = \sqrt{\frac{\sum(y - \hat{y})^2}{n}}$$

$$MAE = \frac{\sum|(y - \hat{y})|}{n}$$

Where:

y- Actual Value of Age
$\hat{y}$- Predicted Value of Age

Classification Problem: The preferred metric for a classification task is Accuracy, which generates probability values against the given ground truth label.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP- True Positive
TN- True Negative

FP- False Positive
FN- False Negative

Confusion Matrix was also used to understand the distribution of accuracy under each age group and each gender.

4. RESULTS AND DISCUSSION

4.1. Ideal duration of Audio

The audio sources in the dataset range in length from 1 to 80 seconds because they are compiled from multiple sources. To predict demographic data like age and gender from speaker audio that generalises across multiple different data sources, the ideal audio length must be determined.

We have chosen CNN+LSTM Based model with MFCC as a feature in which we have used the default number of MFCC as 40 to predict on the same test dataset with default padding and cropping strategy set to left. We have used RMSE and MAE values from the test sets as performance metric.

From the results in the Table 4. 5s chunks perform best in terms of metric. In the case of audio calls from the call centre use case, where the file length exceeds 600 seconds, extracting 5 seconds per speaker would be more than sufficient. This would save computational time an effort and real time response.

Table 4. Audio Length based Performance Assessment

Length of Audio vs Performance			**Custom Test Set**	
Model	**Feature**	**Duration (s)**	**RMSE**	**MAE**
CNN+LSTM	MFCC	3	5.21	3.37
CNN+LSTM	MFCC	5	4.55	2.69
CNN+LSTM	MFCC	10	5.16	3.34
CNN+LSTM	MFCC	15	5.14	3.35

4.2. Optimal Padding and Cropping Strategy

In the case of audio, the length of the 1D sequence must be the same. Our dataset contains variable-length audio sequences, and in real-world use cases, audio will be variable-length as a result of voice activity detection linked speaker diarisation. From the previous experiment we have investigated and understood the effect of audio duration on model performance using RMSE and MAE Metrics. 5s audio length is best suited for our use case.

It is also critical to comprehend the impact of underlying factors that cause the 5s duration to perform better. Padding and cropping are used to make the audio sequence the same length.

Table 5. displays the outcomes of various padding and cropping strategies. It is observed that the Padding of audio on the right and cropping at random yields better results when compared to other strategies for our use case.

Table 5. Padding and Cropping Vs Model Performance

Padding and Cropping Vs Model Performance				Custom Test Set	
Model	**Feature**	**Pad Crop Strategy**	**Duration(s)**	**RMSE**	**MAE**
CNN+LSTM	MFCC	pad - left, crop - random	5	5.21	3.37
CNN+LSTM	MFCC	pad - right, crop - random	5	4.55	2.69
CNN+LSTM	MFCC	pad - left, crop - left	5	5.16	3.34
CNN+LSTM	MFCC	pad - left, crop - right	5	5.14	3.35

4.3. Choosing the Number of MFCC

While using MFCC as the feature extractor for the audio parameter can be set to extract the number of features for every chunk of audio known as Number of MFCC features. By varying this number of MFCC we can also extract differential and acceleration components, also known as Delta and Delta-Delta Components The existing literature makes it clear that including these elements during feature extraction results in better ASR and other speech analytics Model performance.

We worked on varying the values on number of MFCC while keeping all other factors standard to try to find what generalises best for extracting demographic information from the speaker audio. Results from Table 6 verifies the fact that using the additional features of MFCC helps in improving the metric and hence achieving better results on overall custom dataset.

Table 6. Choosing number of MFCC

Selection of number of MFCC		Custom Test Set	
Model	**Hyperparameter (number of MFCC)**	**RMSE**	**MAE**
CNN+LSTM	13	5.49	3.7
CNN+LSTM	40	4.55	2.69

4.4. Experimental Results

4.4.1. CNN Model

The duration of the audio, padding, and cropping method during data ingestion, as well as number of MFCC during feature extraction, have now been identified. We have built models that leverage these values to predict the age, age group and gender information from the speaker audio.

The output from the feature extraction is then passed to the CNN based model followed by self attention. Output of the self attention then predicts all three values age, age group and gender of the input audio. From the Table 7 we can understand that the Wav2Vec feature is giving the best result in terms age RMSE (4.29) and MAE (2.83) for age prediction and for gender prediction the overall accuracy stands at 99.5% with male group having an accuracy of 99.8% and female with 99.1%.

Table 7. CNN Model Result

Feature	Age		Age Group	Gender		
	RMSE	MAE	Accuracy	Male	Female	All
MFCC	4.91	3.15	63.7%	98.8%	99.1%	98.9%
Wav2Vec	4.29	2.83	60.6%	99.8%	99.1%	99.5%

4.4.2. LSTM Model

In case of LSTM model, the input is received from the feature extraction technique selected returns the output. The output is then passed to the self attention layer. The output of self attention is then used to predict age, age group and gender.

From the results in Table 8 We have Wav2Vec Feature giving the best result. In case of LSTM model age group accuracy has increased by over 10% when compared to the CNN based model, with slight decrease in overall performance. We concluded that LSTM based model is able to capture the age group information better than the CNN Model.

Table 8. LSTM Model Result

Feature	Age		Age Group	Gender		
	RMSE	MAE	Accuracy	Male	Female	All
MFCC	5.89	4.16	50.8%	98.9%	97.1%	98.0%
Wav2Vec	4.42	2.66	69.6%	99.6%	99.1%	99.3%

4.4.3. CNN+LSTM Model

CNN Model provides the better result for age and gender prediction while LSTM model captures age group with higher accuracy. In order to improve the complexity and hence to improve the performance we have merged the above said two models CNN and LSTM. This gives our model the ability to learn the features of all the prediction variables better. Model Layer consists of Conv1d layer followed by LSTM and Soft Attention. Self attention then predicts age, age group and gender information.

As per the results in Table 9. MFCC feature is giving better accuracy in terms of age group prediction. In case of the Age and gender prediction the Wav2Vec feature extractor performs better with Age RMSE 4.25 and MAE 2.66, Gender overall accuracy is over 99.3% with female predicted 99.6% and male 99.0% prediction accuracy.

Table 9. CNN+LSTM Model Result

Feature	Age		Age Group	Gender		
	RMSE	MAE	Accuracy	Male	Female	All
MFCC	4.55	2.69	68.3%	99.3%	99.2%	99.2%
Wav2Vec	4.25	2.66	62.0%	99.0%	99.6%	99.3%

4.4.4. Multi CNN+LSTM Model

In case of Multi CNN + LSTM based model we have experimented with 2CNN+LSTM, 3CNN+LSTM results of which are discussed below.

Table 10 provides a relatively better performance in predicting the age and gender with Wav2Vec as feature. The RMSE and MAE values of MFCC feature hasn't improved from the previous model CNN+LSTM.

Table 10. 2CNN+LSTM Model Result

Feature	Age		Age Group	Gender		
	RMSE	MAE	Accuracy	Male	Female	All
MFCC	4.7	2.86	65.9%	99.2%	99.3%	99.2%
Wav2Vec	4.1	2.36	66.6%	99.2%	99.7%	99.5%

Table 11 shows that the results haven't improved much, and the learning has plateaued as the prediction on the test set doesn't provide a significant jump in performance.

We have also experimented with multiple other combinations like CNN + 2 LSTM and 2CNN+2LSTM etc. which doesn't seem to improve on the metrics of the predictive variables and has led us to confirm that the model performance couldn't be improved with variation of CNN and LSTM model.

Table 11. 3CNN+LSTM Model Result

Feature	Age		Age Group	Gender		
	RMSE	MAE	Accuracy	Male	Female	All
MFCC	4.66	2.80	63.8%	99.5%	99.1%	99.3%
Wav2Vec	4.11	2.45	65.9%	99.3%	99.4%	99.4%

4.5. Summary of Results

In Figure 3. results of various models with varying features are shown for Age Prediction. We can see that wav2vec has performed better in terms of RMSE value consistently better across all the model categories.

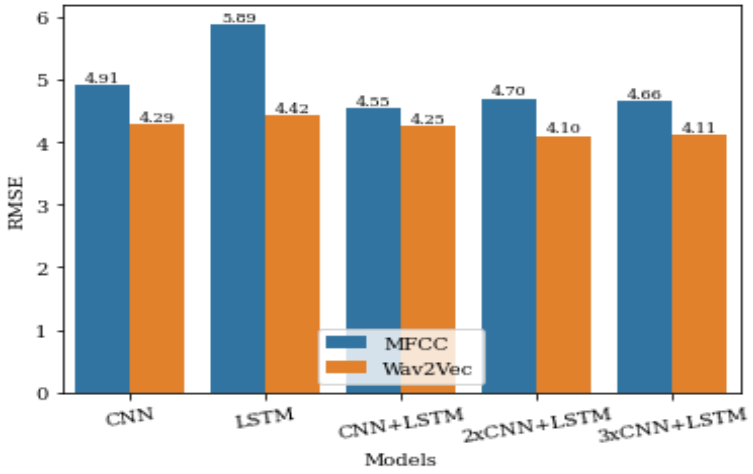

Figure 3. Age Prediction RMSE across models

In Figure 4. We have our models consistently performing better in terms of accuracy of gender prediction in case of wav2vec feature.

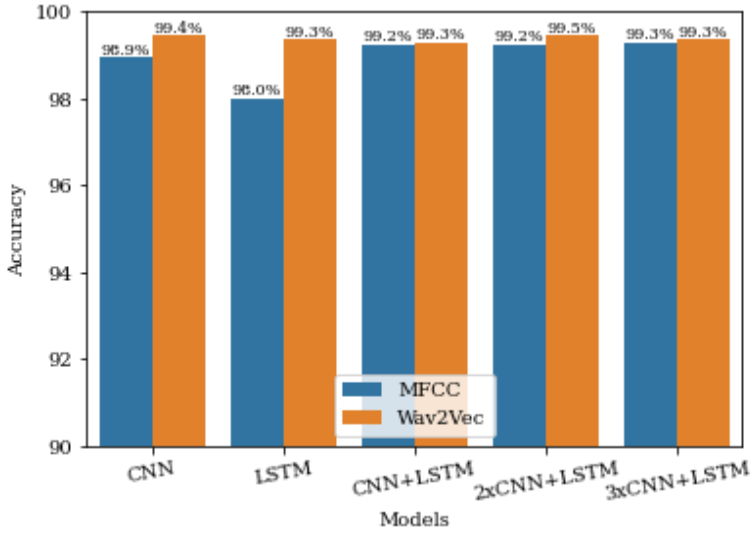

Figure 4. Gender Prediction Accuracy across models

5. CONCLUSION AND FUTURE WORK

In this paper, we have approached the speaker demographic information extraction from speech data by keeping a keen focus on the dataset. A custom dataset with WAV format and 16 KHz sampling rate was curated by selectively combining various standard datasets like TIMIT, NISP, VCTK as well as GMU, thereby improving the robustness of the model being developed to be used in enterprise grade applications. In this paper, we discussed various techniques to improve the model performance by tuning the various stages of model development such as Data Ingestion, Pre-Processing, Feature Extraction, and Model Selection and their effects on model performance. We understand from our research that audio input of a speaker as low as 5 seconds is enough to extract the age, gender and age group of the speaker.

We have explored various approaches with different input features as well as model parameters for age and gender prediction with our custom dataset. Our model capable to work with any speech data across globe from different geolocation. Also the same model can predict both age and gender from speech data. The results showcased that age, age group and gender prediction were best on Wav2vec + 2 CNN + LSTM.

Our future research will be focus on other speech features and wav2vec2 feature as model input features and also will expand the dataset to include more speaker coverages as well as multi language speech data across the globe.

REFERENCES

[1] D. M. Litvinov, "Speech analytics architecture for banking contact centers," in 10th Annual International Scientific and Practical Conference named after AI Kitov Information Technologies and Mathematical Methods in Economics and Management, IT and MM-CEUR Workshop Proceedings, 2021.

[2] S. a. Q. C. Scheidt, "Making a Case for Speech Analytics to Improve Customer Service Quality: Vision, Implementation, and Evaluation.," International Journal of Information Management 45., vol. 45, no. Elsevier: 223–32, 2019.

[3] L. Y, L. Y, D. H and e. al, "Speech databases for mental disorders: A systematic review," General Psychiatry, 2019.

[4] S. Safavi, M. Russell and P. Jančovič, "Automatic Speaker, Age-Group and Gender Identification from Children's Speech," Computer Speech & Language 50., no. Elsevier: 141–56, 2018.

[5] L. Jasuja, A. Rasoo and G. Hajela., "Voice Gender Recognizer Recognition of Gender from Voice Using Deep Neural Networks.," in International Conference on Smart Electronics and Communication (Icosec), 319–24. IEEE., 2020.

[6] M. Buyukyilmaz and A. O. Cibikdiken, "Voice gender recognition using deep learning," in Proceedings of 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016). Paris, France: Atlantis Press, 2016.

[7] T. Maka and P. Dziurzanski, "An Analysis of the Influence of Acoustical Adverse Conditions on Speaker Gender Identification.," in In XXII Annual Pacific Voice Conference (Pvc), 2014.

[8] A. J. Y. C. S. K. Tursunov, "Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module Through Speech Spectrograms.," Sensors 21, 2021.

[9] H. A. Sánchez-Hevia, R. Gil-Pit, M. Utrilla-Manso and M. Rosa-Zurera, "Age Group Classification and Gender Recognition from Speech with Temporal Convolutional Neural Networks," Multimedia Tools and Applications., Vols. Springer, 1–18, 2022.

[10] S. B. Kalluri, D. Vijayasenan, S. Ganapathy and P. Krishnan, "NISP: A Multi-Lingual Multi-Accent Dataset for Speaker Profiling.," ICASSP 2021-2021 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), 2021.

[11] L. F. L. W. M. F. J. G. F. a. D. S. P. J. S. Garofolo, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon technical report, 1993.

[12] J. Yamagishi, C. Veaux and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[13] S. H. Weinberger and S. A. Kunath, "The Speech Accent Archive: towards a typology of English accents.," in In Corpus-based studies in language use, language learning, and language documentation.

[14] Y. A. Wubet and K.-Y. Lian, "A Hybrid Model of Cnn-Svm for Speakers' Gender and Accent Recognition Using English Keywords.," in 2021 Ieee International Conference on Consumer Electronics-Taiwan (Icce-Tw), 2021.

[15] S. B. Kalluri, D. Vijayasenan and S. Ganapathy., "utomatic Speaker Profiling from Short Duration Speech Data.," Speech Communication 121. Elsevier: 16–28, 2020.

[16] S. Schneider, A. Baevski, R. Collobert and M. Auli, "Wav2vec: Unsupervised Pre-Training for Speech Recognition.," arXiv, vol. 1904.05862, 2019.

[17] G. Van Rossum and F. L. Drake., "Python 3 Reference Manual," Scotts Valley, CA: CreateSpace, 2009.

[18] F. William, "The PyTorch Lightning team," Pytorch lightning, 2019.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan and T. Killeen, "PyTorch: An Imperative Style, High-Performance Deep Learning Library.," In Advances in Neural Information Processing Systems 32, 2019.

AUTHORS

Veera Vignesh graduated from Birla Institute of Technology in 2018. Currently, he is a Data Scientist at Accenture Solutions India Pvt Ltd.

Anup Bera has done M.Tech from IIT Kharagpur in 2006 and he has total 16 years of industry experience and currently working at Accenture Solution India Pvt Ltd