Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India

Gender Identification From Children's Speech

Pravin Bhaskar Ramteke *, Amulya A. Dixit †, Sujata Supanekar *, Nagraj V. Dharwadkar †,
and Shashidhar G. Koolagudi *
National Institute of Technology Karnataka, Surathkal, 575025 *
Rajarambapu Institute of Technology,Islampur, Maharashtra †
ramteke0001@gmail.com *, amulya2907@gmail.com †, sujata.supanekar@gmail.com *, nagraj.dharwadkar@ritindia.edu *,
koolagudi@nitk.ac.in †

***Abstract*—Children's speech can be characterized by higher pitch and format frequencies compared to the adult speech. Gender identification task from children's speech is difficult as there is no significant difference in the acoustic properties of male and female child. Here, an attempt has been made to explore the features efficient in discriminating the gender from children's speech. Different combinations of spectral features such as Mel-frequency cepstral coefficients (MFCCs), ΔMFCCs and $\Delta\Delta$MFCCs, Formants, Linear predictive cepstral coefficients (LPCCs); Shimmer and Jitter; Prosodic features like pitch and its statistical variations along with Δpitch related features are explored. Features are evaluated using non linear classifiers namely Artificial Neural Network (ANNs), Deep Neural Network (DNNs) and Random Forest (RF). From the results it is observed that the RF achieves an highest accuracy of 84.79% amongst the other classifiers.**

Index Terms: Gender identification, Pitch, LPCCs, Artificial Neural Network, Deep Neural Network, Random Forest.

I. Introduction

Humans can easily recognize people by hearing to their voice. They are very good at guessing a person's age, gender and emotion from person's voice. The acoustic and linguistic characteristics of child speech are especially different from those of adult speech. Child's speech is observed to have higher pitch value, formant frequencies and specific indistinct articulation with respect to the adult speech [1] [2]. The impact of bandwidth reduction on speech recognition accuracy is greater for children's speech than for adults. The other differentiating parameters includes vocal-tract geometry and less precise control of the articulators. Gender identification of children is difficult than adults, it is confusing to identify whether the speaking child is male or female. Due to underdeveloped vocal tract and thin vocal folds in both male and female child, there is no significant difference in their acoustic-phonetic properties. Identifying the gender of a child speaker based on his/her speech is crucial in telecommunication and speech therapy. Children's at a specific age tend to mispronounce some phonemes which are difficult for them to pronounce. Because of this they replace that difficult phonemes by a other simpler phonemes. This replacement patterns are known as phonological processes [3]. The SLPs analyse children's speech to study the language acquisition patterns in male and female children as they progress in age. In English language, it is observed that the female child acquire the phoneme pronunciation faster compared to male child, it is an object of interest to know the learning pattern in other languages [3]. If these phonological processes persist beyond a certain age they lead to phonological disorders. If the gender identification is done within children voices then the analysis and treatments will become easy.

In this paper, an attempt has been made for the gender identification from children's speech using different combinations of spectral features (MFCCs, Formants and LPCCs); prosodic features (Pitch & its statistical variations) and temporal features (shimmer, jitter). Based on the non-linear nature of the data, Artificial Neural Network (ANNs), Deep Neural Network (DNNs) and Random Forests (RFs) are used for classification tasks.

Rest of the paper is organized as follows. Section II discuss the existing works done in the area of children gender identification. Section III discusses the database used for experimentation. Methodology employed for children gender identification is given in Section IV. Results are discussed in Section V. Paper is concluded by highlighting some future research directions in Section VI.

II. Literature Review

Adult gender identification is easy as compared to the gender identification in children. Many attempts have made for gender identification in adults using various classification approaches and feature combinations. Earlier approaches explored distance measures [4], Gaussian mixture model (GMMs) and hidden Markov model (HMMs) based recognizer [5]. Recent approaches, GMM-Universal Background Model (GMM-UBM) and GMM-Support Vector Machine (GMM-SVM) have shown significant improvement in gender identification. A gender dependent GMM is built using EM training or MAP adaptation on the initial UBM. GMM-SVMs combines the strength of generative GMM-UBM and discriminative power of SVM [6]. Prosodic features are explored for gender classification using dynamic Bayesian networks [7], [8]. Variance normalization and cepstral mean subtraction have been used for gender identification in broad categories of children, young adults, adults and seniors [8]. These approaches mainly focuses on gender identification in adults and do not consider the gender identification in children.

Children's gender identification is a difficult task as there is no significant difference in the acoustic properties of male and female children [9]. Different combinations of spectral,

978-1-5386-6835-1/18/$31.00 ©2018 IEEE

prosodic and excitation source features are explored for the task. Spectral features namely MFCCs, prosodic features such as pitch are mostly used in many approaches towards this task. As the features are already known the classification task requires the implementation of classifiers on the data. Study over 21 frequency sub-bands regions of the spectrum show that the frequency range less than 1.8 kHz and greater than 3.8 kHz are efficient in discriminating gender in older children [10]. Frequencies greater than 1.4 kHz are useful for the youngest children [10]. The openSMILE feature set, a combination of spectral features such as MFCCs, log mel-frequency band, line spectral pairs and prosodic features like F0 along with its statistical variations, shimmer and jitter have shown importance in children gender identification [11]. As of today, very few approaches have focused on gender identification from children's speech. One of the attempts include the use of GMM-UBM and GMM-SVM systems [10]. In this, the age-dependent and age-independent analysis is done.[10]. Both GMM-UBM and GMM-SVM are implemented on different age group criteria and then the performance is examined [10].

GMMs & SVMs are most commonly used in gender identification with MFCCs, formants, pitch and its statistical variations [8]. However, there is scope for increasing the accuracy using different combinations of features and implementing suitable classifiers based on the nature of the data. Based on the shortcomings of literature, ANNs, DNNs and RFs are used for classification tasks. MFCC and pitch are common and well known in gender identification area. Shimmer, jitters and LPCCs are some acceptably useful features for speech tasks, hence they are considered for children gender identification results [12].

III. Database Used

The database used in this work is CMU Kids Corpus, which consists of sentences read aloud by children both male and female in English language [13]. The database was originally designed to create a training set of children's speech for the SPHINX II automatic speech recognizer under the LISTEN project at Carnegie Mellon University (CMU). There are total of 818 audio records. The children range in age from 6 years to 11 years. The female records contain 544 samples where male records have 274 samples.

IV. Methodology

The proposed framework for the identification of children's gender is shown in Fig. 1. The approach is divided into three stages. The first stage involves pre-processing the speech signal. In pre-processing, the silences and unvoiced regions in the speech signal are removed. Second stage is feature extraction. The voiced regions are considered for the feature extraction where the features efficient in gender classification such as Mel-frequency Cepstral Coefficients (MFCCs), Linear predictive cepstral coefficients (LPCCs), Formants, Pitch, Shimmer and Jitter are extracted. In the last step, the efficiency of various combinations of the extracted features is evaluated using different classifiers namely Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs) and Random Forest (RFs).

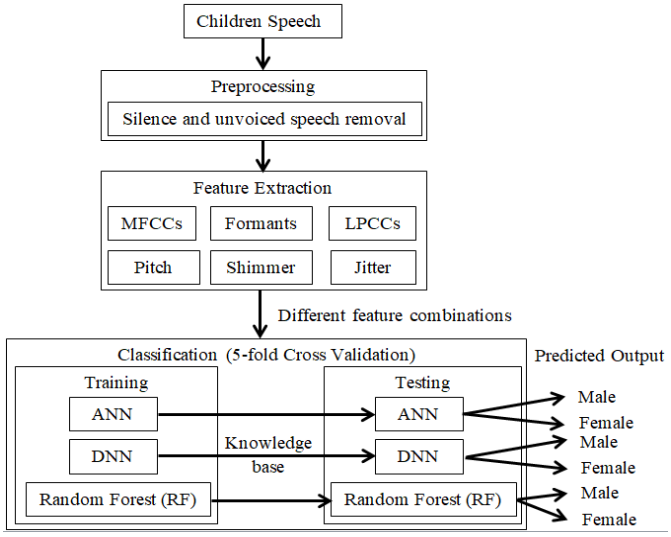

Fig. 1. Flow diagram of the proposed approach

A. Silence and unvoiced speech removal

Most of the gender related information lies in the voiced region of speech. The speech recordings of children consists of many silence and unvoiced regions. The silence and unvoiced speech regions are removed from the speech using short time energy feature. Low energy is observed in unvoiced & silence regions whereas voiced regions are characterized by high energy values.

$$E_T = \sum_{n=1}^{N} s^2(n) \tag{1}$$

where, E_T is the energy of T^{th} frame. N is the length of frame (number of samples in a frame). The threshold is set based on the average energy (avg_energy) value. It can be represented using,

$$thr\_avg\_energy = a * avg\_energy \tag{2}$$

where, a is constant may vary from 0 to 1. thr_avg_energy represents threshold value for the segmentation. From the analysis threshold value is set 0.15. The energy values below threshold are considered as either silence or unvoiced and these frames are removed.

B. Feature Extraction

Features efficient in discriminating female and male voice in children speech should be identified. Effectiveness of 6 features (MFCCs, LPCCs, Formants, pitch, shimmer & jitter) and their combinations are considered for the evaluation.

1) Mel-frequency Cepstral Coefficients (MFCCs): The most commonly used acoustic features in gender classification are MFCCs [14]. They play a significant role in applications such as speech recognition, speaker recognition, etc. MFCCs mimics human speech production and speech perception, by logarithmic perception of loudness [14]. Total 39 MFCC features are extracted from speech signal which include 13 MFCCs, 13 ΔMFCCs and 13 $\Delta\Delta$MFCCs respectively.

2) Formants: Formants frequencies change with different vocal tract configurations corresponding to different resonance [15]. The difference in formant frequencies of adult male and female can be observed [16]. As the vocal tract length increase the values of formant frequencies get reduced [15]. Children have higher formants frequencies than both female and male adults. The formant extraction is done using LPC analysis method [17]. Four formants are considered for the classification task.

3) Pitch: Pitch is the rate of vocal fold vibration also known as the fundamental frequency of speech signal [18]. Typically, adult male pitch values ranges from 85-155 Hz. For female, it varies between 165Hz to 255Hz; for children the approximate range is 200Hz to 350 Hz. Use of pitch may give a clue to gender classification from children speech. The pitch contour is extracted from speech signal using probabilistic YIN (PYIN) algorithm [18], modified autocorrelation method for pitch estimation. Here, pitch along with its statistical variations is considered. First order derivative (Δpitch) of pitch is also used for the gender identification task.

4) Shimmer and Jitter: Jitter refers to the variability of fundamental frequency [19]. It is mainly affected by lack of control of vocal fold vibration. Shimmer is affected because of reduction in tension of vocal folds [19]. Absolute and relative values are extracted for both shimmer and jitter. Process of extraction of shimmer and jitter is given in [19]. In adults, the values of absolute jitter are found to be larger in males as compared to females. On the other hand, the values of relative jitter are larger in females. Inorder to evaluate the same in children gender, shimmer and jitter are considered for the analysis.

5) Linear predictive cepstral coefficients (LPCCs): LPC are the coefficients of an auto-regressive model of a speech frame [20]. The all-pole representation of the vocal tract transfer function is as given by

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{k=1}^{n} a_k . z^{-k}} \tag{3}$$

where a_p are the prediction coefficients and G is the gain. LPCCs are obtained directly using [20],

$$LPCC_i = \sum_{k=1}^{i-1} (k - i/i) LPCC_{i-k} a_k \tag{4}$$

LPCCs are well known for their performance in many speech related tasks such as speech recognition, speaker recognition, etc. Hence it is considered for the analysis.

C. Classification

The classification task involves the implementation of various classifiers for gender identification task. The classifiers are chosen mainly based on the non-linear nature of data. An Artificial Neural Network (ANNs), Deep Neural Network (DNNs) and ensemble method random forest (RFs) are used to develop the classification model.

1) Artificial Neural Network (ANNs): An Artificial Neural Network is a computational model based on the structure and functions of biological neural networks. Neurons are organized in layers: input layer, hidden layer and output layer. These different neuron layers may perform different types of transformations on their given inputs [21]. The layer between input and output layers is called hidden layer. The model is trained by adjusting the weights of the neuron for classification task [22]. Feed forward neural network is considered for the experimentation. The number of hidden neurons are set equal to the the mean of the neurons in the input and output layers. From the analysis activation function is set to 'sigmoid'.

2) Deep Neural Network (DNN): DNN is a classifier based on feedforward artificial neural networks [23]. The architecture of the deep neural network is shown in Fig. 2. The architecture contains input layer, output layer and hidden layers. The input layer contains neurons which depend on the number of feature combinations. The output layer has the neuron representing the outcome of classification. The architecture consists of multiple hidden layers with non-linear activation functions [21] [24]. Each node in a layer uses of same non-linear activation function. Commonly used activation functions are 'ReLu', 'tanh', 'sigmoid' and 'softmax'. Combinations of these activation functions are implemented using Deep Neural Network Algorithm [23] [24]. The number of hidden layers and corresponding number of neurons are set as suggested in [24].

From the analysis, it is observed that, three hidden layers are sufficient to evaluate the performance on the considered size of dataset. The non-linear activation function 'ReLu' is set for the hidden layer. The activation functions 'sigmoid' is set for the output layer. Sigmoid activation function have shown good accuracy for binary classification as it is primarily used for binary classification problems unlike the 'softmax' function which is mostly used for multi-class classification. The impact of increase in number of efficient features has been observed in the hidden layer structure. The number of neurons has to be set properly in hidden layers as the number of features increases to achieve good accuracy. Table III gives the details of essential parameters, namely number of neurons in input layer, number of neurons in each hidden layer, activation functions set for hidden layers and output layers set for different feature combinations of features. The batch size set for training this classifier is 10. The DNN classifier deals with all these different combinations of parameters resulting in achieving good results. However, different combinations of hidden layers and activation functions has to be implemented to achieve high accuracy results.

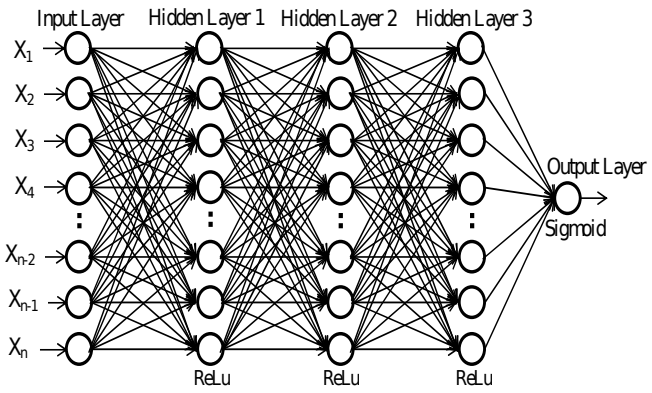

Fig. 2. Architecture of Deep neural network

3) Random Forest (RFs): The RF classifier is formed by combining multiple tree based classifiers where each tree is built from a random vector independently sampled from the total input vector [25]. The classifier uses bagging, a method to generate a training set by arbitrarily drawing the replacement from the training dataset. This is done for each feature combination considered. Class label is assigned to a test sample by taking the most popular class voted by all the tree predictors of the forests [25]. In the case of random forest, with the progression in forest building, it tries to overcome the internal unbiased generalization error, hence efficient in estimation of missing data [25]. This enables the model to inherit an ability to achieve good accuracy even when the large proportion of the dataset is unrecoverable and unbalanced class population.

V. Results and Discussion

In children speech, there is no significant difference in the characteristics of male and female children, as their vocal tract is undeveloped and have similar size & length, vocal folds are thin which results in high pitch value. This increase the difficulty in classification task. It is difficult for a human to distinguish the male and female child from their speech. It indicates that the features extracted for the classification task may be highly non-linear in nature. CMU Kid Corpus is considered for the analysis, where recordings of female child contain 544 speech samples and male child recording have 274 speech samples. Spectral features such as MFCCs, formants and LPCCs, prosodic features like pitch & its statistical variations, jitter and shimmer measurements are considered for the task. Baseline system is developed using 39 MFCC features. Further, different combinations of the features are tested to evaluate the performance of gender identification system as shown in Table I. The classifiers efficient in discriminating the data having non-linear nature are considered for the experimentation; namely Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs) and Random Forest (RF). Essential parameter set for the DNNs are given in the Table III along with the classification accuracy for each parameter set. 80% of the instances are used for training and 20% for testing with 5-fold cross validation. Each classifier is trained for every combination of the feature set and accuracy is evaluated. The accuracy is percentage of total number of instances correctly identified. Table II shows the average accuracy of classification for various combinations of features.

TABLE I
Features and their combinations considered for gender identification

Sr. No.	Number of Features	Feature Combinations
1	39	MFCC (13) + ΔMFCC (13) + ΔΔMFCC (13)
2	43	MFCC (13) + ΔMFCC (13) + ΔΔMFCC (13) + Pitch (4)
3	47	MFCC (13) + ΔMFCC (13) + ΔΔMFCC (13) + Pitch (4) + ΔPitch (4)
4	51	MFCC (13) + ΔMFCC (13) + ΔΔMFCC (13) + Pitch (4) + ΔPitch (4) + Formant (4)
5	55	MFCC (13) + ΔMFCC (13) + ΔΔMFCC (13) + Pitch (4) + ΔPitch (4) + Formant (4) + Shimmer (2) + Jitter (2)
6	68	MFCC (13) + ΔMFCC (13) + ΔΔMFCC (13) + Pitch (4) + ΔPitch (4) + Formant (4) + Shimmer (2) + jitter (2) + LPCC (13)

MFCCs are observed to be efficient in many speech tasks [14], hence MFCCs are considered as baseline features to evaluate the baseline performance of each classifier. ANN trained using 39 MFCC features, achieve an average accuracy of 72.30%. DNN is observed to achieve an average accuracy of 71.66%. Random forest achieve an average accuracy of 84.21%. As the size of the dataset is small, DNN may not be efficient for the task, hence observed to achieve least accuracy compared to the other two classifiers (For the details of the parameters set for the DNN classifier for different feature combinations refer Table III). RF is efficient in building an accurate classifier which can efficiently run on the small and large sized datasets of non-linear nature. Hence observed to achieve good accuracy compared to the other two classifiers [25]. Adult speech can be easily discriminated by observing pitch values. Here, an attempt has been made to evaluate the role of pitch in discriminating gender in children. Pitch along with it statistical variations : minimum, maximum & standard deviations are considered along with baseline features. The performance of ANN & DNN is improved by 0.60% and 0.57% receptively compared to the baseline ANN & DNN system. Whereas the performance of Random Forest is observed to improve by 0.18%. Though the performance of all three classification systems are not improved much, this shows that children pitch has gender specific information. Pitch derivative are also evaluated for the same task. ANN, DNN & RF are trained using MFCCs (39), pitch (4) and pitch derivatives (4) (feature vector of size: 47). It is observed that there is no significant influence of the pitch derivative on the performance of all three systems. The average accuracy achieved is given in Table II.

Formants represents the resonance of vocal tract. In adults, there is a small deviation in formant values of same sound unit for male and female [16]. The same is attempt to evaluate

TABLE II
AVERAGE CLASSIFICATION ACCURACY OF MALE AND FEMALE CHILDREN GENDER

Sl. No.	Classifier Used	Number of Features Considered					
		39	43	47	51	55	68
1	ANN	72.30%	72.90%	72.80%	72.80%	74.10%	76.20%
2	DNN	71.66%	72.23%	72.32%	73.59%	78.11%	78.25%
3	Random Forest	84.21%	84.39%	84.30%	84.39%	84.49%	84.79%

TABLE III
ACCURACIES ACHIEVED THROUGH DNN USING DIFFERENT HIDDEN LAYERS AND ACTIVATION FUNCTIONS

Sl. No.	No. of Features	No. of Neurons Set in Each Layer DNN					Activation Function for Each Layer of DNN				Accuracy Achieved (%)
		Input	L1	L2	L3	Output	L1	L2	L3	Output	
1	39	39	39	39	39	1	ReLu	ReLu	ReLu	sigmoid	71.66
2	43	43	43	43	43	1	ReLu	ReLu	ReLu	sigmoid	72.23
3	47	47	47	47	47	1	ReLu	ReLu	ReLu	sigmoid	72.32
4	51	51	51	51	51	1	ReLu	ReLu	ReLu	sigmoid	73.59
5	55	55	55	55	55	1	ReLu	ReLu	ReLu	sigmoid	78.11
6	68	68	68	68	68	1	ReLu	ReLu	ReLu	sigmoid	78.25

in children speech. 4 formants are used with MFCCs (39), pitch (4) and pitch derivatives (4) [feature vector of size 51] to train the classifiers considered. The performance of ANN is improved by 0.50%. DNN is observed to improve the performance of the gender classification by 1.93% compared to the baseline system. Whereas the performance of Random Forest is observed to improve by 0.18%. This shows that the formants play important role in children gender identification. Shimmer and jitter are the measure of cyclic variation in speech. It is observed to be efficient in speaker identification. Absolute & relative jitter and shimmer are considered for the analysis along with the 51 features as shown in Table I. Results shows that, there is a significant improvement in the results, when shimmer and jitter are used in combination with MFCCs (39), pitch (4), pitch derivatives (4) and formants (4). ANN achieve an improvement of 2.00%. DNN is observed to achieve an improvement of 6.45% compared to the baseline system, where as Random forest achieve an improvement of 0.28%. The accuracy of classification is further observed to improve by adding 13 LPCC features to the previous feature vector. LPCCs are observed to be efficient in many speech tasks such as speech recognition, emotion recognition, etc. Here with the use of 13 LPCCs, the performance of ANN, DNN, & Random Forest is improved by 3.9%, 6.59%, 0.58% respectively. This shows that the shimmer, jitter and LPCCs are efficient in discriminating the gender in children. From the results, it can be observed that, though the features are observed to improve the performance of each classifier using combination of features, the performance of ANN and DNN is very poor compared to the Random forest. Though ANN & DNN are efficient in modelling the non-linear data, small size of may affect the performance of ANN and DNN as they need large data for training. Random forest are efficient in discriminating features non-linear in nature. It also work well with the small sized data. Random forest outperform ANN & DNN with highest average accuracy of 84.79% for feature vector size of 68. By looking at the performance

TABLE IV
RESULTS OF PREVIOUS RESEARCH WORK DONE BY OTHER RESEARCHERS

Sl. No.	Classifiers used	Database used	Features used	Accuracy
1	GMM-UBM [10]	OGI Kids Corpus	MFCC (19)+ ΔMFCC (19) + ΔΔ MFCC (19)	78.53%
2	GMM-SVM [10]	OGI Kids Corpus	MFCC (19)+ ΔMFCC (19) + ΔΔ MFCC (19)	84.14%

improved with LPCCs, RF is trained using MFCCs (39), pitch (4) & LPCCs (13). The average accuracy achieved is 84.77, it is equivalent to the performance of RF trained using 68 (refer Table I). Hence, MFCCs (39), pitch (4) & LPCCs (13) are observed to be sufficient for the gender identification in children using Random Forest classifier.

The previous research work too identified the gender from children's speech [10]. This approach used the OGI Kids Corpus Database with three different age groups namely 5-9 years, 9-13 years and 13-16 years. On whole dataset, the highest accuracy achieved using age independent GMM-UBM is 67.39%, where as age dependent GMM-UBM achieves an accuracy of 71.76%. When performance of age dependent GMM-UBM is evaluated for each age group, the highest accuracy of 78.53% is achieved for age group 13-16 years (refer Table IV). Whereas for GMM-SVM based classifier, the performance of age independent GMM-SVM is 77.44% is achieved on complete dataset. The age dependent GMM-SVM achieve an accuracy of 79.18%. The performance of age dependent GMM-SVM is evaluated separately on each age group shows that, the highest accuracy of 84.14% is achieved for age group 9-13 years. As we used the CMU Kids Corpus Database, we have children voices with 6 to 11 age range. The proposed approach use entire dataset evaluation and do not divide into any sort of categories. The state-of-the-art approaches are implemented on different datasets, hence it is

difficult to compare them with our approach. The analysis of the results shows that the performance of the proposed system is better, as the accuracy achieved is 84.79% using Random Forest.

VI. Summary and Conclusion

The task of gender identification from children's speech is difficult compared to adult gender identification. Features used for this task are MFCCs (39), Pitch (4), ΔPitch (4), Formant (4), Shimmer (2), Jitter (2) and LPCCs (13). To evaluate the efficiency of the proposed approach different combinations of these features are used for classification. Based on the non-linear nature of the data, classifiers efficient in discriminating non-linear data namely, ANN, DNN & Random Forest are considered. The random forest classifier outperform the other classifiers with an average accuracy of 84.79% for gender classification. DNN achieves an average accuracy of 78.25%. ANN did not gave satisfying accuracy results as compared with other two classifiers. Random forests are non-linear classifiers and they work very well on small data especially on binary classification problems [1]. Further, the performance of the classification can be improved by using combination of spectral, prosodic and excitation source features. Spectral features extracted from sub-bands regions can be considered with proposed set of features, as the spectrum show that the frequency range less than 1.8 kHz and greater than 3.8 kHz are efficient in discriminating older children [10]. Frequencies greater than 1.4 kHz are useful for the youngest children [10]. Prosodic features such as statistical variations of pitch, e.g. minimum, maximum, mode, median, quartile of pitch etc. can also be considered for the classification task. Also it is possible to classify the children's by their age and then evaluate the accuracy for CMU kid corpus.

References

[1] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proceeding of Speech Prosody*, 2016, pp. 84–88.

[2] R. Martins, I. Trancoso, A. Abad, and H. Meinedo, "Detection of childrens voices," *Proceedings of the I Iberian SLTech*, pp. 77–80, 2009.

[3] D. Ingram, *Phonological disability in children*. London:Edward Arnold, 1976.

[4] K. Wu and D. G. Childers, "Gender recognition from speech. part I: Coarse analysis," *The journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[5] E. S. Parris and M. J. Carey, "Language independent gender identification," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 685–688.

[6] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1605–1608.

[7] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. G. Bauer *et al.*, "Comparison of four approaches to age and gender recognition for telephone applications," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1089.

[8] M. Li, C.-S. Jung, and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 2826–2829.

[9] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[10] S. Safavi, M. Russell, and P. Jančovič, "Identification of age-group from children's speech by computers and humans," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 243–247.

[11] H. Kaya, A. A. Salah, A. Karpov, O. Frolova, A. Grigorev, and E. Lyakso, "Emotion, age and gender classification in childrens speech by humans and machines," *Computer Speech & Language*, vol. 46, pp. 268–283, 2017.

[12] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[13] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids speech corpus," *Corpus of children's read speech digitized and transcribed on two CD-ROMs, with assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania*, 1997.

[14] V. Tiwari, "MFCC and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[15] J. N. Holmes, W. J. Holmes, and P. N. Garner, "Using formant frequencies in speech recognition," in *Fifth European Conference on Speech Communication and Technology*, 1997, pp. 2083–2086.

[16] A. P. Simpson, "Phonetic differences between male and female speech," *Language and Linguistics Compass*, vol. 3, no. 2, pp. 621–640, 2009.

[17] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129–134, 1993.

[18] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 659–663.

[19] M. Farrús and J. Hernando, "Using jitter and shimmer in speaker verification," *IET Signal Processing*, vol. 3, no. 4, pp. 247–257, 2009.

[20] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[21] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, vol. 23, no. 3, pp. 325–350, 2017.

[22] P. H. Sydenham and R. Thorn, *Handbook of measuring system design*. Wiley Online Library, 2005, vol. 3.

[23] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Italian Computational Linguistics Conference (CLiC-it)*, 2014.

[24] G. Panchal, A. Ganatra, Y. Kosta, and D. Panchal, "Review on methods of selecting number of hidden nodes in artificial neural network," *International Journal of Computer Theory and Engineering*, vol. 3, no. 2, pp. 332–337, 2011.

[25] L. Breiman, "Random Forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.