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

A Stacked Technique for Gender Recognition through Voice

Pramit Gupta
Computer Science Department
Jaypee Institute of Information Technology
Noida, India

Somya Goel
Computer Science Department
Jaypee Institute of Information Technology
Noida, India

Archana Purwar
Computer Science Department
Jaypee Institute of Information Technology
Noida, India

Abstract**—***Detecting the gender of a person (male or female) through their voice seems to be a very trivial task for humans. Our minds are trained over the course of time to detect the differences in voices of males and females. Our ears work as the front end, receiving the audio signals which our brain processes and makes the decision. But it is a challenging problem for computers. Gender classification has applications like, it is able to improve the intelligence of a surveillance system, analyze the customer's demands for store management, and allow the robots to perceive gender etc. This paper proposes a stacked machine learning algorithm to determine gender using the acoustic parameters of voice sample and compares its performance with existing classifiers as CART, Random forest and neural network.*

Keywords—Machine learning; voice; gender; neural network

I. Introduction

Gender of a person plays a significant role in our day to day interactions with computers and society. As gender possesses distinguished information concerning social activities, automatic gender classification is receiving increasing attention [1].Gender classification is a trivial problem for a person as over time human brain learns how to classify voice into that of male and female. Many spoken-dialog systems work only on the sequences of words from speaker's voice, they miss utilizing the other useful information that can be inferred from speech such as gender, dialect, emotion and age. Gender recognition has various potential uses in different application such as :

a) Surveillance and Security Systems : Classifying gender can contribute towards increasing intelligence of security and surveillance systems by assisting in the investigation of criminals who intentionally try to hide their identity information [2]. It would also help evaluate gender specific threat level [3].

b) Video Games and Mobile Application : In games, male and female players have different preferences; automatic gender classification would provide their preferred game characters or contents. For example, character features, such as gait, can be analyzed using gender classification techniques [4]. Applying different gait patterns in virtual characters according to gender will improve the sense of reality of the game [5]. In multimedia apps, this is used to facilitate the user by tailoring applications according to user's gender [19].

c) Human-Computer Interaction : In the area of HCI [6] like virtual assistants, computers are needed to identify and verify gender in order to enhance performance based on personalized information. Gender classification can provide customized services to users by adapting according to their gender [7].

This paper proposes an approach for gender recognition using stacked machine learning algorithms based on their speech using 20 acoustic features and compares with existing machine learning algorithms.

II. Related Work

Speech processing based several types of research works have been continuing from a few decades ago as a field of digital signal processing (DSP). A gender detection system was constructed by extracting first formant and pitch using linear predictive analysis [8]. The research concluded that both features are higher in female's voice than that of males. To determine gender, K-nearest neighbor classification was used, but with some constant k value it lacks comparing similar outputs. Whereas modern deep learning algorithms overcome these limitations because they provide more flexibility as shown in [9, 10]. Chen et al. worked on cepstral peak prominence, spectral magnitude and harmonic to noise ratio [11]. The research results varied with the age group. While detecting gender, they achieved 60% accuracy in 8-10 years and 94% accuracy in 16-17 years group of children.

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

Most of the work were using pitch factor. In paper [12], energy entropy, short time energy and zero crossing rates were given as an input to fuzzy logic and neural network individually and calculated percentage of presence of male and female features in speech sample from both the systems. The proposed technique was tested on Harvard-Haskins database [13] and achieved 65% accuracy, higher than 50% and 60% attained by fuzzy logic and neural network individually. In [14], researchers computed mean values for three features energy entropy, zero crossing rate and short time energy from training dataset and used genetic approach to calculate percentage of gender identified. They further calculated and compared specificity, sensitivity and achieved accuracy of 79%. This paper aims at improving the accuracy of gender detection using proposed methodology.

III. Proposed Methodology

This section proposes a stacked machine learning algorithm for gender detection system as shown in Figure 1. To apply stacked machine learning algorithm, dataset was divided into 3 portions, 2 of which were used for training purposes and one for final testing. We trained CART (Classification and Regression Trees), SVM (Support Vector Machine) and Neural Network models over one portion of the training data, upon evaluation of models we noticed that for a given input, different models made different predictions. Different models use different techniques to fit the data and hence work on different underlying properties. The predictions of different models shows that there are instances where the outcomes of different models differ from each other and the ground truth result. We cannot trust a single model to give accurate results. Typically majority voting is performed in stacked methods. The label predicted by majority of models is given as the final result, but in some cases even majority voting does not ensure that ground truth is predicted.

In order to give a stronger prediction we used ensemble learning technique which takes multiple predictions derived from different models as its input and train another model (Random Forest / SVM / Neural Network) over these combined results to predict final result as depicted in Figure 1. Second proportion of the training data was used as test data for SVM, CART and NN and the predictions made were used to train the stacked model. Stacking refines predicted results and hence reflects increase in performance compared to that of single trained model.

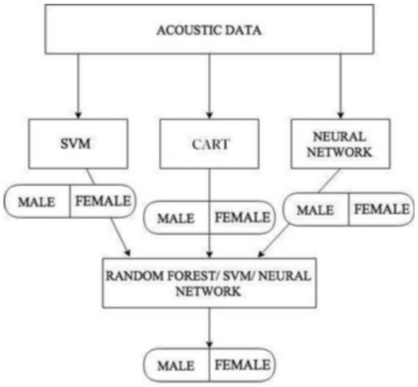

Figure 1. Stacked Model Representation

IV. Experiments and Results

A. Dataset

Different voice samples have different set of acoustic properties. We collected acoustic data from Harvard-Haskins Database of Regularly-Timed Speech [13]; recording three males and three females native English speakers spoken syllables, from VoxForge [15] ; a collection of 1200 speech samples of male and female speakers, from Festvox CMU_ARCTIC Speech Database at Carnegie Mellon University [16] . From above data we were able to gather around 3160 recorded samples male and female voices varying within different ranges of frequencies. We extracted acoustic audio features using the tools provided by warbleR and seewave library in R. These outputs are saved into CSV file where twenty columns depicting features and one extra column as label for gender (male or female).

B. Experiment Setup

The dataset is divided into two parts. First part is used to train CART, SVM and Neural Network models. Radial Biased kernel was used in training SVM. Neural Network implementation had dense layers with RELU activation followed by last layer with softmax activation. It was trained using binary cross entropy loss and Adam optimizer. Weights were initialized as mentioned in [17]. The prediction of these models on second part of training data along with their ground truth labels is used as the dataset to train an ensemble of Random Forest, SVM and Neural Network model. Majority voting amongst these models provides the final prediction. The third portion of the dataset is used for testing of all individual models and ensemble model

C. Results

The models were tested on a dataset containing 950 samples of which 475 were male and 475 were female.

Stacked models marginally outperform others. Amongst the stacked models, neural network shows the most gain. Accuracy increases by 2% from the base model. Precision shows when the model predicts male; how often it is actually male and specificity signifies how often the model predicts female when the ground truth is actually female. F-score is weighted average of recall and precision. Precision gets a boost of 2.82%, specificity increases by 3% and F-score gains 2%. All the above values are tabulated in Table I.

V. Conclusion and Future Work

In this paper, we developed an acoustic-based computer program solution to determine gender from speech. In designing a gender recognition system, the feature selection is one of the most important factors. Some papers focused on finding single best feature or approach to determine gender, but a single feature was not enough to classify gender. Papers using fuzzy logic turn out to be a flexible study as true or false decisions are dependent on membership functions. It was already studied that learning algorithms like neural network give more accuracy than fuzzy logic based system. Thus we designed a system for considering twenty acoustic factors together and applied three models. Since every model predicted a different result as per their tree structure or hyper parameters we set, stacked model used those results to train new models which reduced error. This improved the efficiency by re-evaluating and re-assuring the predicted results.

We achieved 96.74% accuracy with stacked Random Forest and stacked Neural Network model. Stacking SVM on the other hand did not produce any changes. Factors which are related to energy and frequency domain could also be clubbed for better classification. As the problem is related to voice samples; with more diverse and bigger dataset, accuracies attained may decrease and also we could have acknowledged the error reduction improvement more prominently.

I. Performance Metrics Table

Model	Accuracy	Error Rate	Specificity	Precision	Recall	F-score
CART	95.05	4.94	93.89	94.03	96.21	95.10
Neural Network	95.57	4.42	96.84	96.76	94.31	95.52
SVM	95.78	4.21	95.78	95.78	95.78	95.78
Stacked SVM	96.01	4.00	95.78	95.80	96.21	96.00
Stacked RF	96.73	3.26	96.84	96.83	96.63	96.72
Stacked NN	97.05	2.94	96.84	96.85	97.26	97.05

References

[1] Udry J. R., "The nature of gender," *Demography,* vol. 31, pp. 561-573, 1994.

[2] Demirkus M., Garg K., and Guler S., "Automated person categorization for video surveillance using soft biometrics," in *Biometric Technology for Human Identification VII*, 2010, p. 76670P.

[3] Jain A. K., Ross A., and Prabhakar S., "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology,* vol. 14, pp. 4-20, 2004.

[4] Paluchamy M., Suganya D., and Ellammal S., "Human gait based gender classification using various transformation techniques," *IJRCCT,* vol. 2, pp. 1315-1321, 2013.

[5] Gnanasivam P. and Muttan S., "Gender classification using ear biometrics," in *Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012)*, 2013, pp. 137-148.

[6] Amayeh G., Bebis G., and Nicolescu M., "Gender classification from hand shape," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, 2008, pp. 1-7.

[7] Hoffmeyer-Zlotnik J. H. and Wolf C., *Advances in cross-national comparison: A European working book for demographic and socio-economic variables*: Springer Science & Business Media, 2003.

[8] Rakesh K., Dutta S., and Shama K., "Gender Recognition using speech processing techniques in LABVIEW," *International Journal of Advances in Engineering & Technology,* vol. 1, pp. 51-63, 2011.

[9] Khan S. A., Ahmad M., Nazir M., and Riaz N., "A comparative analysis of gender classification techniques," *Middle-East Journal of Scientific Research,* vol. 20, pp. 1-13, 2014.

[10] Mäkinen E. and Raisamo R., "An experimental comparison of gender classification methods," *pattern recognition letters,* vol. 29, pp. 1544-1556, 2008.

[11] Chen G., Feng X., Shue Y.-L., and Alwan A., "On using voice source measures in automatic gender classification of children's speech," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] Meena K., Subramaniam K. R., and Gomathy M., "Gender classification in speech recognition using fuzzy logic and neural network," *Int. Arab J. Inf. Technol.,* vol. 10, pp. 477-485, 2013.

[13] The Harvard-Haskins Database of Regularly-Timed Speech, http://www.nsi.edu/~ani/download.html

[14] Jayasankar T., Vinothkumar K., and Vijayaselvi A., "Automatic Gender Identification in Speech Recognition by Genetic Algorithm," *Appl. Math,* vol. 11, pp. 907-913, 2017..

[15]oxForge Speech Corpus, http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/8kHz_16bit/.

[16] Festvox CMU_ARCTIC Speech, http://festvox.org/cmu_arcti.

[17] He K., Zhang X., Ren S., and Sun J., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.