International Journal of Innovative Research in Engineering
Volume 3, Issue 3 (May-June 2022), PP: 649-654.
www.theijire.com
ISSN No: 2582-8746

Voice Based Gender Recognition Using Deep Learning

Dr. Sayyada Fahmeeda[1], Mohamed Abdullah Ayan[2], Mohamed Shamsuddin[3], Aliya Amreen[4]
[1]Assistant Prof. Dept. of Computer Science and Engineering, PDA College Of Engineering, Kalaburagi. Karnataka, India.
[2,3,4] Dept. of Computer Science and Engineering, PDA College Of Engineering, Kalaburagi. Karnataka, India.

How to cite this paper
Dr.Sayyada Fahmeeda[1],Mohamed Abdullah Ayan[2],Mohamed Shamsuddin[3],Aliya Amreen[4], "Voice Based Gender Recognition Using Deep Learning", IJIRE-V3I03-649-654.

Abstract: *The present paper describes the development in voice based gender recognition. Speech recognition has various applications including human to machine interaction, sorting of telephone calls by gender categorization, video categorization with tagging and so on. Currently, Deep learning is a popular trend which has been widely utilized.in various fields and applications, exploiting the recent development in digital technologies and the advantage of storage capabilities from electronic media. Recently, research focuses on the combination of ensemble learning techniques with the semi-supervised learning framework aiming to build more accurate classifiers. The paper aim on gender recognition by voice utilizing a new ensemble semi-supervised convolution neural network algorithm. Our preliminary numerical experiments demonstrate the classification efficiency of the proposed algorithm in terms of accuracy, leading to the development of stable and robust predictive models.*

Keywords*: speech recognition, semi supervised learning, classification.*

I.INTRODUCTION

One of the most common means of communication in the world is through voice. In the real world, it is possible for a person to verify the gender of a person through voice. Voice is filled with lots of linguistic features. These voice features are considered as the voice prints to recognize the gender of a speaker. The recorded voice is considered as the input to the system, which then the system process to get voice features .Examine the input and compare it with the trained model, carry out calculations based on the algorithm used and gives the latest matching output.

The human voice is the most compatible medium for interaction among human beings. When sound comes out from a vocal throat it carries much regional, bio-logical, and surrounding atmospheric data. Using those kinds of information, we can find out human language, gender, age, and accent, emotional and present state. Gender recognition is a technique to determine gender categories by the speaker's voice signal analysis. Gender recognition is a technique which is often utilized to determine the gender category of a speaker by processing speech signals.

Speech signals taken from a recorded speech can be used to acquire acoustic attributes such as duration, intensity, frequency and filtering. As gender recognition is a significant task, it increases the performance of applications such as human–computer interaction, emotion recognition, person identification, sorting of telephone calls by gender categorization, online advertisements involving voice. The proposed approach recognize the gender from audio datasets and also live databy focusing on gender recognition by voice utilizing a new ensemble semi-supervised self-labeled algorithm. The data set can be trained with different machine learning algorithms. Our preliminary numerical experiments demonstrate the classification efficiency of the semi-supervised self-labeled algorithm in terms of accuracy, leading to the development of stable and robust predictive models.

II. OBJECTIVE

The voice of human speech is an effective communication method consisting of unique semantic linguistic and para-linguistic features such as gender, age, language, accent, and emotional state. The sound waves consisting of human voice are unique among all creatures producing sound since every single wave carries a different frequency. Identifying human gender based on voice has been a challenging task for voice and sound analysts who deploy numerousapplications. There are a set of features used for recognizing the voice gender. Among the most common features utilized for voice gender recognition are Mel scaled power spectrogram (Mel), Mel- frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma), spectral contrast (Contrast), and tonal centroid features (Tonnetz). By getting the extracted features combined with the gender label as a form of a training set, ML techniques are used to build a high-quality model for recognizing the voice gender.

III.METHODOLOGY

The dataset used for detecting gender from the audio files is retrieved from kaggle, which is a free speech corpus and acoustic model repository for open source speech engines. It is a large-scale collection of voices of both genders. From the collected audio files the powerful discriminating features are extracted with which a CSV file is created. With this CSV file various models are trained using Support Vector Machine, Decision Trees, Gradient Tree Boosting, Random forests, and accuracy is calculated. The goal is to compare outputs of various models and suggest the best model that can be used for gender recognition by voice in real-world inputs.

First, the voice (.wav files) is converted to a form that the system can understand. Preprocessing needs to be done on the file to avoid external noises. After the noises have been removed the feature extraction process can be carried out. It is much necessary to find out the acoustic features that have a high discriminative power to classify the genders.

Next step, is to train the machine with collected features from the dataset to make themachine capable to classify the genders of the voice. Here the machines are trained with four algorithms Support Vector Machine, Decision tree, Gradient Tree Boosting, and Random Forest.

Calculate the accuracy of each algorithm with the dataset.After finding out the best algorithm over the dataset (based on the accuracy), a particular algorithm will be used to find out the gender of the real world input given while testing.Mel-Spectrogram. Mel-spectrogram computes a Mel- scaled power spectrogram coefficient. An object of Mel- spectrogram type represents an acoustic time-frequency representation of sound, as shown in Figure. The power spectral density P (f, t) is sampled into a number of points around equally spaced times $t_i$and frequencies f_j (on a Mel- frequency scale).

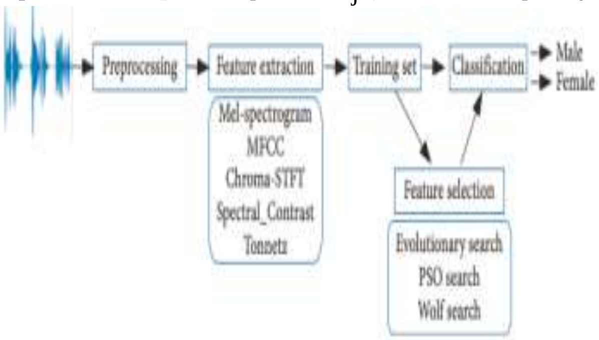


Figure (1): MFCC Block Diagram

MFCC represents accurately the vocal tract that is a filtered shape of a human voice and also manifests itself in the envelope of a short-time power spectrum, as shown in Figure. In order to compute MFCCs, a set of sequential steps should be followed:

(1) Framing the Signal into Short Frames. The audio signal is framed into 20–40 ms (25 ms is standard) frames to over- come changes in the sample in a short time period as it is constantly changed in a long period of time.

(2) Periodogram of Power Spectrum. This calculates for each frame the periodogram estimation of the power spectrum, which identifies the frequencies in the frame.

(3) Applying the Mel Filterbank to the Power Spectra (or Summing the Energy in Each Filter). A filter is required for estimating the energies in various frequency regions that appear in a group of aggregated periodogram bins because of unnecessary information inperiodogram spectral estimation. Hence, the Mel filter bank estimates the energy near 0 Hz and then for higher frequencies as there is less concern for variations.

(4) Logarithm of All Filterbank Energies. Large variations of energies are scaled using a logarithmic scale as there are no different sounds in large energies. The logarithmic scale is

(5) DCT of the Log Filterbank Energies. Because of the correlation in filterbank energies that lead to overlapping, the DCT is used to decorrelate the energies. This generates diagonal covariance matrices as features.

(6) 2-13 DCT Coefficients. Higher DCT coefficients are chosen to reduce the fast changes in the filterbank energies and discard the rest.

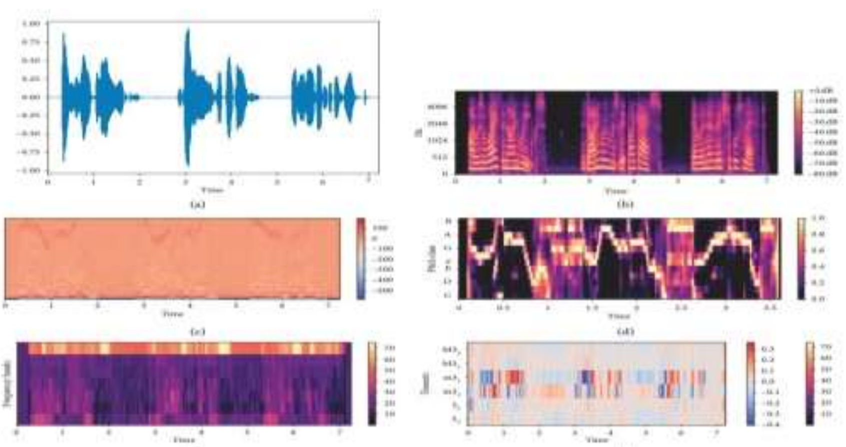


Figure (2): A British English female's voice and its features: (a) Voice sample; (b) mel-spectogram; (c) MFCC; (d) chromagrams; (e) spectral contrast; (f) tonal centroids (Tonnetz)

Deep Gender Recognition (DGR)

The proposed methodology for speech gender classification includes a set of stages as briefly discussed below. The stages start by converting the voice, from its abstract representation, into a consistent form in order to extract the relevant features. Then, the relevant features are selected as inputs for building a classifier model for recognizing the gender of a human voice. In addition, a DL model is being built to automatically extract useful features and feed them into a fully connected artificial neural network (ANN) for classification. However, here, a set of process for extracting features for other models rather than DL and classification techniques are summarized as follows.

1. Voice Pre-processing. A transmitted voice is inevitably vulnerable to noise interference and voice attenuation that needs a pre-processing process to purify it for feature ex- traction. This phase shows a set of steps as follows.

A/D Signal Conversion. A/D signal conversion is used to convert the given voice from the analogue to the digital signal by common sampling and quantization techniques. The A/D conversion formulates the signal in an un- derstandable form by machine for easy manipulation.

2. Preemphasis Process. Because of attenuation at high- frequency segments of the voice signal, there is a necessary need to use a preemphasis filter. The preemphasis filter flattens the signal (or speech) waveforms. The process filters low-frequency interference, especially power frequency interference at low-frequency segments, and emphasizes the high-frequency portions in order to produce a high-pass filter to carry out spectral analysis interference. This process occurs after A/D conversion by the first-order digital pre- emphasis filter equation coefficient with the value ranging commonly within.

3. Frame Blocking and Hamming Window. The frame blocking is a process of handling the filtered digital signal into a number of N small frame segments with adjacent frames separated by M (M<N). The process of the Hamming window minimizes speech signal discontinuities before and after each frame within the window frame. This method is popularly used in the MFCC before the Mel-Frequency warping step where Mel scales are calculated. The analytical representation of the Hamming window is given by where w(n) is the window operation, *n* is the number ofindividual samples, and *N* is the total number of speech samples.

4. Fast Fourier Transform (FFT). The FFT algorithm is in general used for estimating the discrete Fourier transform (DFT) of any sequence, or its inverse form. In the speech voice signal, the FFT converts each frame of those N samples from the time-domain signal into a form of frequency domain The FFT is considered a computationally signal. Hence, a pre-processing phase is needed to prepare the speech signals as an input for a set of feature extraction techniques. These sets of features and a voice gender as a label represent the training set for building a classifier model in order to recognize the voice speech gender. For visualization, Figure shows a voice sample, which is a British English female's voice and its features. The features used in this paper are as follows.

2.2.1. Mel-Spectrogram. Mel-spectrogram computes a Mel- scaled power spectrogram coefficient. An object of Mel-spectrogram type represents an acoustic time-frequency representation of sound, as shown in Figure .The power spectral density P(f, t) is sampled into a number of points around equally spaced times t_i and frequencies f_j (on a Mel- frequency scale). MFCC. MFCC represents accurately the vocal tract that is a filtered shape of a human voice and also manifests itself in the envelope of a short-time power spectrum.

(1) Framing the Signal into Short Frames. The audio signal is framed into 20–40 ms (25 ms is standard) frames to over- come changes in the sample in a short time period as it is constantly changed in a long period of time.

(2) Period gram of Power Spectrum. This calculates for each frame the period gram estimation of the power spectrum, which identifies the frequencies in the frame.

(3) Applying the Mel Filter bank to the Power Spectra (or Summing the Energy in Each Filter). A filter is required for estimating the energies in various frequency regions that appear in a group of aggregated period gram bins because of unnecessary information in period gram spectral estimation. Hence, the Mel filter bank estimates the energy near 0 Hz and then for higher frequencies as there is less concern for variations.

IV. EXPERIMENTAL RESULTS

A set of experiments are conducted for evaluating the contributions, which include studying the efficiency of extracted features, evaluating different learning techniques, and analyzing the three natural optimizers used for feature selection. This section also shows the datasets, experimental parameters, and settings and then presents the evaluation of the presented contributions.

Experimental Settings. A standard dataset of artificial voices from the study in is used. The dataset consists of 20 languages. Each language has 16 voice samples of eight files for each gender. The artificial voice is a signal mathematically produced for regenerating the time and spectral characteristics of the human speech. These artificial voices have bandwidth between 100 Hz and 8 kHz, which significantly affects the performance of linear and nonlinear.

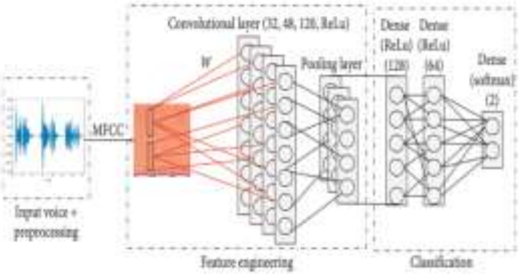


Figure (3): One-Dimensional conventional neural network

Telecommunication systems. The artificial voice is mainly used for objective evaluation of speech processing systems and devices. A single channel with continuous activity (i.e., without pauses) is sufficient for measuring characteristics. The advantage of generating artificial voice is that it is more easily generated and has smaller variability than real voice.

Voice Feature Effect and Correlation. In order to study what features are relevant to build an optimal classifier, the correlation between features has to be examined to demon- strate how they are related to each other. Four feature types, in the present work, are considered including MFCCs, Chroma, Mel, and Tonnetz. The correlation between features is presented in Figure that shows a scattered plot representing the correlation relationship among different feature types. Each chart contains a linear regression equation that formulates the evaluated feature values. In addition, it clarifies the R^2 correlation coefficient. R^2 is a statistical measure that determines how close the real data points are fitted by the linear regression model. This means that if the R^2 value is close to 1, the data are highly fitted to the regression line, and there is no difference in their effects on tested labels, or there is, in contrast, a bad correlation to the labels.

In particular, as shown in Figure 4, the best R^2 of 0.332 is between the MFCC and Chroma features. In contrast, based on the chart, the worst correlation occurs between the Chroma and Contrast features with R^2 equal to 0.35. At each feature category, the MFCC feature has the best correlation with the Chroma feature with R^2 equal to 0.332 and worst correlation with the Tonnetz feature with R^2 equal to 0.0012. The Chroma feature has the worst correlation with the Mel feature with R^2 equal to 0.0004 compared to the other feature categories. The Mel feature has a worse correlation value in comparison with the Tonnetz feature.

In summary, the MFCC, Chroma, and Mel features show an efficient performance as they are more related to each other. The reason is that these features extract high-energy coefficients from the signal where, in contrast, the other features concern about the tone of the musical signals. This answers the research question RQ1 that ensures the im- portance of selecting more suitable features for possibly building more accurate classifiers for voice paralinguistic information aspects.

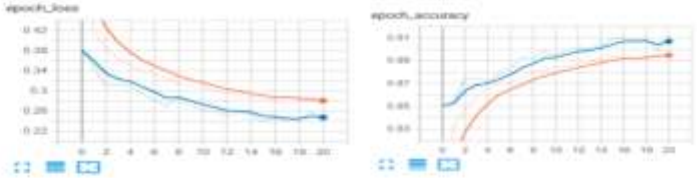
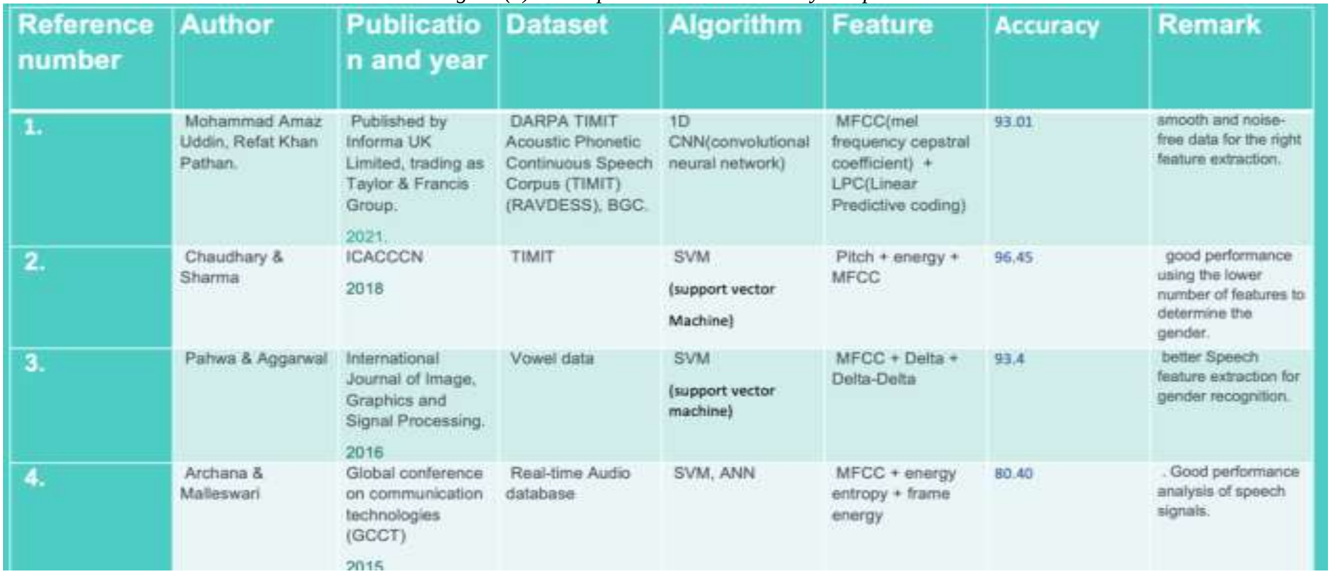


Figure (4): The Epoch Loss and Accuracy Graph



Reference number	Author	Publicatio n and year	Dataset	Algorithm	Feature	Accuracy	Remark
1.	Mohammad Amaz Uddin, Refat Khan Pathan.	Published by Informa UK Limited, trading as Taylor & Francis Group. 2021.	DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus (TIMIT) (RAVDESS), BGC.	1D CNN(convolutional neural network)	MFCC(mel frequency cepstral coefficient) + LPC(Linear Predictive coding)	93.01	smooth and noise-free data for the right feature extraction.
2.	Chaudhary & Sharma	ICACCCN 2018	TIMIT	SVM (support vector Machine)	Pitch + energy + MFCC	96.45	good performance using the lower number of features to determine the gender.
3.	Pahwa & Aggarwal	International Journal of Image, Graphics and Signal Processing. 2016	Vowel data	SVM (support vector machine)	MFCC + Delta + Delta-Delta	93.4	better Speech feature extraction for gender recognition.
4.	Archana & Malleswari	Global conference on communication technologies (GCCT) 2015	Real-time Audio database	SVM, ANN	MFCC + energy entropy + frame energy	80.40	. Good performance analysis of speech signals.

Figure (5): Literature Survey Review

V. CONCLUSION

Hence the model obtained show us that we can use acoustic properties of the voices and speech to detect the voice gender. From the literature survey we make out thatthere are various algorithms used by different surveyor and we have taken into account their accuracy .we have studied the structure of their algorithm used and decided to make our own voice gender model with more accuracy. Our primary aim is to develop a model which is more accurate than the previously developed one.

Gender identification is one of the major problems in the area of signal processing. The system deals with finding the gender of a person using vocal features. Some of the human vocal features reveals the fact that classifying gender goes beyond the frequency and pitch of a person. One of the most challenging problems faced is feature selection from wide range of features, which is discriminating factor in classifying the gender of a person. The proposed method deals with deep learning techniques to recognize gender based on voice.

Hence we decided to use two algorithms Gaussian and Tensor flow framework using recurrent neural network .In the further preceding will implement the steps to build a model and check the accuracy on gender recognition by voice utilizing a new ensemble semi-supervised convolution neural network algorithm. Our preliminary numerical experiments demonstrate the classification efficiency of the proposed algorithm in terms of accuracy, leading to the development of stable and robust predictive models.

References

1. *Alsulaiman, M., Ali, Z., & Muhammad, G. (2011, November). Gender classification with voice intensity. In 2011 UKSim 5th European symposium on computer modeling and simulation (pp. 205–209). IEEE. https://doi.org/10.1109/EMS.2011.37*
2. *Archana, G. S., & Malleswari, M. (2015, April). Gender identification and performance analysis of speech signals. In 2015 Global conference on communication technologies (GCCT) (pp. 483–489). IEEE. https://doi.org/10.1109/GCCT.2015.7342709*
3. *Chaudhary, S., & Sharma, D. K. (2018, October). Gender identification based on voice signal characteristics. In 2018 International conference on advances in computing, communication control and networking (ICACCCN) (pp. 869–874). IEEE https://doi.org/10.1109/ICACCCN.2018.8748676*
4. *Chen, G., Feng, X., Shue, Y. L., & Alwan, A. (2010, September). On using voice source measures in automatic gender classification of children's speech [Paper presentation]. Eleventh Annual Conference of the International Speech Communication Association, pp. 26–30.*
5. *Djemili, R., Bourouba, H., & Korba, M. C. A. (2012, May). A speech signal based gender identification system using four classifiers. In 2012 International conference on multimedia computing and systems (pp. 184–187). IEEE. https://doi.org/10.1109/ICMCS.2012.6320122*
6. *Ertam, F. (2019). An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, 156, 351–358. https://doi.org/10.1016/j.apacoust.2019.07.033*
7. *Gaikwad, S., Gawali, B., & Mehrotra, S. C. (2012). Gender identification using SVM with combination of MFCC. Advances in Computational Research, 4(1), 69–73. Ghosal, A., & Dutta, S. (2014, February).*
8. *Automatic male-female voice discrimination. In 2014 International conference on issues and challenges in intelligent computing techniques (ICICT) (pp. 731–735). IEEE. https://doi.org/10.1109/ICICICT.2014.6781371*
9. *Holzinger, A. (2019). Introduction to Machine learning & Knowledge Extraction (MAKE). Machine Learning and Knowledge Extraction, 1(1), 1–20. https://doi.org/10.3390/make1010001*
10. *Keyvanrad, M. A., & Homayounpour, M. M. (2010, May). Improvement on automatic speaker gender identification using classifier fusion. In 2010 18th Iranian conference on electrical engineering (pp. 538–541). IEEE. https://doi.org/10.1109/IRANIANCEE.2010.5507010*
11. *Kim, H. S. (n.d.). Linear predictive coding is all-pole resonance modeling. Center for Computer Research in Music and Acoustics, Stanford University. Linear Interpolation. (2021, June 25). In Wikipedia. https://en.wikipedia.org/wiki/Linear_- interpolationLivieris, I.*
12. *E., Pintelas, E., & Pintelas, P. (2019). Gender recognition by voice using an improved self-labeled algorithm. Machine Learning and Knowledge Extraction, 1(1), 492–503. https://doi.org/10. 3390/make1010030*
13. *Madhu, N. (2009). Note on measures for spectral flatness. Electronics Letters, 45(23), 1195–1196. https://doi.org/10.1049/el.2009.1977*
14. *Majkowski, A., Kołodziej, M., Pyszczak, J., Tarnowski, P., & Rak, R. J. (2019, September). Identification of gender based on speech signal. In 2019 IEEE 20th International conference on computational problems of electrical engineering (CPEE) (pp. 1–4). IEEE. https://doi.org/10.1109/CPEE47179. 2019.8949078*
15. *Pahwa, A., & Aggarwal, G. (2016). Speech feature extraction for gender recognition. International Journal of Image, Graphics and Signal Processing, 8(9), 17. https://doi.org/10.5815/IJIGSP.2015. 09.03*