Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Machine Learning based Gender Voice Recognition

Umarani Kunsoth, Saresh Kumar Ellamla

Department of Electronics and Communication Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

Abstract

Gender Recognition (GR) means recognizing the gender of the person whether the person is men or women. It is a significant task for human beings, as many communal functions precariously rely upon correct gender awareness. Automatic human GR by machines has currently received symbolic attention in computer vision community. Features like face, 3D body shape, gait (manner of walking), footwear, fundamental frequency of voice etc are used for gender recognition. The ability to do automatic recognition of human gender is essential for several systems that process or exploit human-source information. Common examples are information retrieval, human computer, or human-robot intercommunication. The result of an AGR system can be used for achieving meta-data information useful for annotating audio files. Moreover, gender is an important cue that can be exploited for improving intelligibility of man-machine interaction, or just, for decreasing the investigation space in applications such as speaker recognition or surveillance systems. Gender recognition has great importance in forensics, games, business intelligence, demographic survey, visual surveillance. It is completely known that human's speech encloses linguistic content, identity as well as the emotion of speaker.

Keywords: Gender recognition, machine learning, Automatic human GR.

1. Introduction

Language is the main way for people to communicate. In addition to the message meaning contained in language, it also contains the transmission of emotions. Through emotions, tone, and other messages; even if the other party does not understand the meaning of the message in the language, one can still feel the speaker's emotions in words. In recent years, the use of artificial intelligence and deep learning for emotion recognition has attracted much attention. The industrial applicability of emotion recognition is quite comprehensive and has good development potential. In various applications in daily life, human–computer interaction has gradually been replaced by voice operations and dialogues from touch-sensitive interfaces. Speech recognition is widely used in transportation, catering, customer service systems, personal health care, and leisure entertainment.

The Bangla language is well-known around the world, and it is the fifth most spoken language on the planet. The population of Bangladesh speak two different varieties of Bangla. Few people speak the local language of their region. The mainstream Bangla language, which is spoken by about 290 million people, is another variety. There are 55 regional languages spoken in Bangladesh's 64 districts. A regional language, also known as a dialect, is a language a child learns organically without the use of written grammar, and that varies by region. It is a characteristic of languages that are widely spoken in a given location that causes morphological differences in the sounds of the ideal language or literary language. Despite regional variations, the Bangla language can be divided into six classes: Bangla, Manbhumi, Varendri, Rachi, Rangpuri, and Sundarbani. This study primarily focused on seven regional languages; Khulna, Bogra, Rangpur, Sylhet, Chittagong, Noakhali, and Mymensingh divisions, which all belong to one of these classes, and one was chosen at random. A person's regional language can be identified by the wave frequency (pronunciation) of a word pronounced in Bangla.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

The machine learning algorithms attempt to learn from these deep neural networks by extracting specific features and information. Prior to 2006, searching deep architecture inputs was not a predictable straight forward task; however, the development of deep learning algorithms helped resolve this issue and simplified the process of searching the parameter pace of deep architectures. Deep learning models can also operate as a greedy layerwise unsupervised pre-training. This means that it will learn hierarchy from extracted features from each layer at a time. Feature learning is achieved by training each layer with an unsupervised learning algorithm, which takes the features extracted from the previous layer and uses it as an input for the next layer. Thus, feature learning will attempt to learn the transformation of the previously learned features at each new layer. Each iteration feature learning adds one layer of weights to a deep neural network. The resulted layers with learned weights can eventually be loaded to initialize a deep supervised predictor. Using deep architectures has proven to be more efficient in representing non-linear functions in comparison to shallower architectures. Studies have shown that fewer parameters are required to represent a certain non-linear function in a deep architecture in comparison with the large number of parameters needed to represent the same function in a shallower architecture. This shows that deeper architectures are more efficient from a statistical point of view. Chen et. al [1] examines the use of application software for ideological and political theory courses in colleges and universities and the application of application software for ideological courses and analyzes the use of apps in the management of ideological and political theory courses in colleges and universities. Including the development and design of ideological and political theory applications, combining deep learning and CTC algorithms to build acoustic models, using server-client interaction, designing an ideological and political theory course app based on speech recognition and deep learning, and forming an offline speech recognition system software platform, the app provides a certain reference for improving the skills of teaching managers.

Jayne et. al [2] proposes an automated recognition of the geographical origin and gender of a voice sample based on the six regional dialects of the United Kingdom. Twenty-six features are extracted from 17,877 voice samples and then used to design, implement and evaluate machine learning classifiers based on Artificial Neural Networks (ANNs), Support Vector Machine (SVM), Random Forest (RF) and k-nearest neighbors (k-NN) algorithms. The results suggest that the proposed approach could be applicable for areas such as e-commerce and the service industry, and it provides a contribution to NLP audio research.

Wani et. al [3] proposed a time-frequency method for the classification of gender-based on the speech signal. Various techniques like framing, Fast Fourier Transform (FFT), auto-correlation, filtering, power calculations, speech frequency analysis, and feature extraction and formation are applied on speech samples. The classification is done based on features derived from the frequency and time domain processing using the Support Vector Machines (SVM) algorithm. SVM is trained on two speech databases Berlin Emo-DB and IITKGP-SEHSC, in which a total of 400 speech samples are evaluated. An accuracy of 83% and 81% for IITKGP-SEHSC and Berlin Emo-DB have been observed, respectively.

2. Literature survey

Yadav et. al [4] considered standard data set for gender identification using machine learning approach. Various approaches are considered to achieve greater accuracy and low error (specificity/MCC). Number of experiments are done to learn optimal weights for different value of K = 5, 10, 15 and 20. From range of experiment, it is clear that higher K fold validation is giving better accuracy all time.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Jha et. al [5] provided a speech emotion recognition framework that is both reliable and efficient enough to work in real-time environments. Speech emotion recognition can be performed using linguistic as well as paralinguistic aspects of speech; this work focusses on the latter, using non-lexical or paralinguistic attributes of speech like pitch, intensity and mel-frequency cepstral coefficients to train supervised machine learning models for emotion recognition. A combination of prosodic and spectral features is used for experimental analysis and classification is performed using algorithms like Gaussian Naïve Bayes, Random Forest, k-Nearest Neighbours, Support Vector Machine and Multilayer Perceptron.

Altalbe et. al [6] proposes a deep learning method based on long-term short-term memory (LSTM) that can be used with preprocessing, segmentation, and retrieval of audio signals from the GTZAN dataset. The simulation results show that the proposed algorithm can effectively improve the audio fingerprint-based data retrieval accuracy and overcome traditional methods' drawbacks. Compared with existing methods, the proposed LSTM method has achieved good results. The precision, recall, accuracy and F-measure of LSTM is 96.54%, 96.15%, 98.56% and 0.96% respectively. In the real world, the recommended audio fingerprint recognition system effectively works through voice applications, especially in heterogeneous portable consumer devices or online audio distributed systems.

Hourri et. al [7] proposing a new way to use deep neural networks (DNNs) in speaker recognition, in the purpose to facilitate to DNN to learn features distribution. They have been motivated by our previous work, where they have proposed a novel scoring method that works perfectly with clean speech, but it needs improvements under noisy conditions. Moreover, this new method outperformed both i-vector/PLDA and our baseline system in both clean and noisy conditions.

Vaijayanthi et. al [8] proposes a synthesis approach to combine the Mel Frequency Cepstral Coefficients (MFCC) with the vibration rate (PITCH) in order to characterize the emotion according to its respective vocal speech signals. The RAVDESS dataset is utilized here and the extracted features are modelled using the K-Nearest Negibhour and Decision Tree classifier for recognizing the eight emotions.

Xue et. al [9] proposed attention-based two-pathway Densely Connected Convolutional Networks (ATP-DenseNet) is proposed to identify the gender of handwriting. There are two pathways in ATP-DenseNet: Feature pyramid could extract hierarchical page feature, and attention-based DenseNet (A-DenseNet) could extract the word feature by fusing Convolutional Block Attention Module (CBAM) and dense connected block.

Methodology

Therefore, this project aims to develop an effective gender recognition by human voice using fast fourier transform-based machine learning framework for enhanced accuracy as compared to existing MFCC, linear predictive coding techniques.

3. Proposed system

Fast Fourier Series

A fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. The DFT is obtained by decomposing a sequence of values into components of different frequencies. This operation is useful in many fields, but computing it directly from the definition is often too slow to be practical. An FFT

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

rapidly computes such transformations by factorizing the DFT matrix into a product of sparse (mostly zero) factors.[2] As a result, it manages to reduce the complexity of computing the DFT from {\textstyle O\left(N^{2}\right)}{\text style O\left(N^{2}\right)}, which arises if one simply applies the definition of DFT, to {\textstyle O(N\log N)}{\textstyle O(N\log N)}, where {\displaystyle N}N is the data size. The difference in speed can be enormous, especially for long data sets where N may be in the thousands or millions. In the presence of round-off error, many FFT algorithms are much more accurate than evaluating the DFT definition directly or indirectly. There are many different FFT algorithms based on a wide range of published theories, from simple complex-number arithmetic to group theory and number theory.

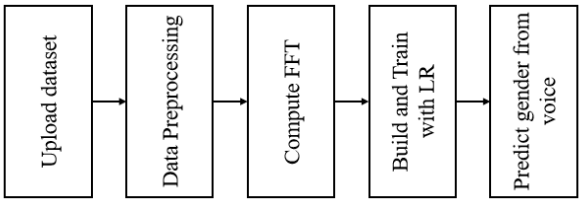


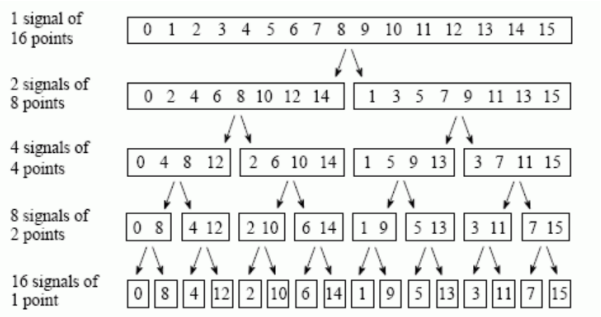
Fig. 1: Block diagram of proposed system.

How the FFT works

The FFT is a complicated algorithm, and its details are usually left to those that specialize in such things. This section describes the general operation of the FFT, but skirts a key issue: the use of complex numbers. If you have a background in complex mathematics, you can read between the lines to understand the true nature of the algorithm. Don't worry if the details elude you; few scientists and engineers that use the FFT could write the program from scratch.

In complex notation, the time and frequency domains each contain one signal made up of N complex points. Each of these complex points is composed of two numbers, the real part and the imaginary part. For example, when we talk about complex sample X, it refers to the combination of ReX and ImX. In other words, each complex variable holds two numbers. When two complex variables are multiplied, the four individual components must be combined to form the two components of the product. The following discussion on "How the FFT works" uses this jargon of complex notation. That is, the singular terms: signal, point, sample, and value, refer to the combination of the real part and the imaginary part.

The FFT operates by decomposing an N point time domain signal into N time domain signals each composed of a single point. The second step is to calculate the N frequency spectra corresponding to these N time domain signals. Lastly, the N spectra are synthesized into a single frequency spectrum.



ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Fig. 2: The FFT decomposition. An *N* point signal is decomposed into *N* signals each containing a signal point. Each stage uses an *interlace decomposition,* separating the even and odd numbered samples.

Sample numbers in normal order		Sample numbers after bit reversal	
Decimal	*Binary*	*Decimal*	*Binary*
0	0000	0	0000
1	0001	8	1000
2	0010	4	0100
3	0011	12	1100
4	0100	2	0010
5	0101	10	1010
6	0110	6	0100
7	0111	14	1110
8	1000	1	0001
9	1001	9	1001
10	1010	5	0101
11	1011	13	1101
12	1100	3	0011
13	1101	11	1011
14	1110	7	0111
15	1111	15	1111

Fig. 3: The FFT bit reversal sorting. The FFT time domain decomposition can be implemented by sorting the samples according to bit reversed order.

separate stages. The first stage breaks the 16-point signal into two signals each consisting of 8 points. The second stage decomposes the data into four signals of 4 points. This pattern continues until there are N signals composed of a single point. An interlaced decomposition is used each time a signal is broken in two, that is, the signal is separated into its even and odd numbered samples. The best way to understand this is by inspecting Fig. 3.3 until you grasp the pattern. There are Log2N stages required in this decomposition, i.e., a 16-point signal (24) requires 4 stages, a 512-point signal (27) requires 7 stages, a 4096-point signal (212) requires 12 stages, etc. Remember this value, Log2N; it will be referenced many times in this chapter.

Now that you understand the structure of the decomposition, it can be greatly simplified. The decomposition is nothing more than a reordering of the samples in the signal. Fig 3.4 shows the rearrangement pattern required. On the left, the sample numbers of the original signal are listed along with their binary equivalents. On the right, the rearranged sample numbers are listed, also along with their binary equivalents. The important idea is that the binary numbers are the reversals of each other. For example, sample 3 (0011) is exchanged with sample number 12 (1100). Likewise, sample number 14 (1110) is swapped with sample number 7 (0111), and so forth. The FFT time domain decomposition is usually carried out by a bit reversal sorting algorithm. This involves rearranging the order of the N time domain samples by counting in binary with the bits flipped left-for-right.

The next step in the FFT algorithm is to find the frequency spectra of the 1-point time domain signals. Nothing could be easier; the frequency spectrum of a 1-point signal is equal to itself. This means that nothing is required to do this step. Although there is no work involved, don't forget that each of the 1-point signals is now a frequency spectrum, and not a time domain signal.

The last step in the FFT is to combine the N frequency spectra in the exact reverse order that the time domain decomposition took place. This is where the algorithm gets messy. Unfortunately, the bit reversal shortcut is not applicable, and we must go back one stage at a time. In the first stage, 16 frequency spectra (1 point each) are synthesized into 8 frequency spectra (2 points each). In the

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

second stage, the 8 frequency spectra (2 points each) are synthesized into 4 frequency spectra (4 points each), and so on. The last stage results in the output of the FFT, a 16-point frequency spectrum.

In below figure shows how two frequency spectra, each composed of 4 points, are combined into a single frequency spectrum of 8 points. This synthesis must undo the interlaced decomposition done in the time domain. In other words, the frequency domain operation must correspond to the time domain procedure of combining two 4-point signals by interlacing. Consider two-time domain signals, abcd and efgh. An 8-point time domain signal can be formed by two steps: dilute each 4-point signal with zeros to make it an

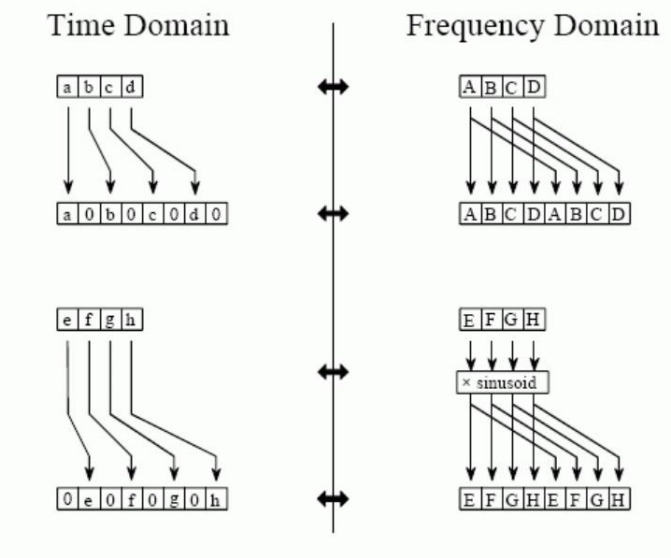


Fig. 4: The FFT synthesis. When a time domain signal is diluted with zeros, the frequency domain is duplicated. If the time domain signal is also shifted by one sample during the dilution, the spectrum will additionally be multiplied by a sinusoid.

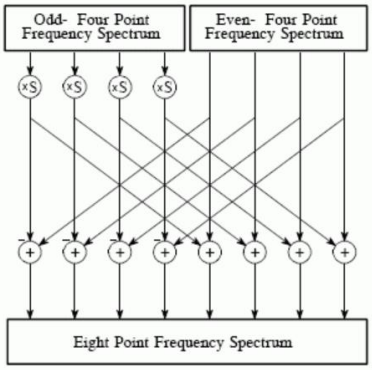


Fig. 5: FFT synthesis flow diagram. This shows method of combining two 4-point frequency spectra into a single 8-point frequency spectrum. The xS operation means that the signal is multiplied by a sinusoid with an appropriately selected frequency.

8-point signal, and then add the signals together. That is, abcd becomes a0b0c0d0, and efgh becomes 0e0f0g0h. Adding these two 8-point signals produces aebfcgdh. Diluting the time domain with zeros corresponds to a duplication of the frequency spectrum. Therefore, the frequency spectra are combined in the FFT by duplicating them, and then adding the duplicated spectra together.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

In order to match up when added, the two-time domain signals are diluted with zeros in a slightly different way. In one signal, the odd points are zero, while in the other signal, the even points are zero. In other words, one of the time domain signals is shifted to the right by one sample. This time domain shift corresponds to multiplying the spectrum by a sinusoid. To see this, recall that a shift in the time domain is equivalent to convolving the signal with a shifted delta function. This multiplies the signal's spectrum with the spectrum of the shifted delta function. The spectrum of a shifted delta function is a sinusoid.

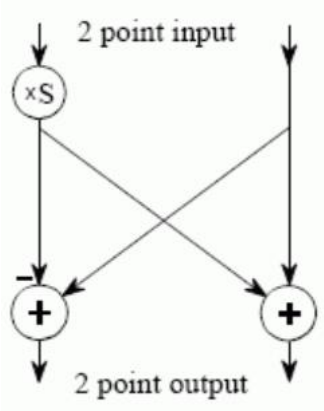


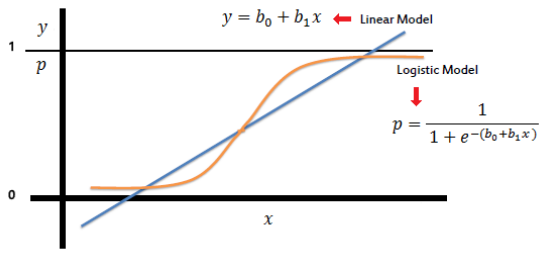
Fig. 6: The FFT butterfly. This is the basic calculation element in the FFT, taking two complex points and converting them into two other complex points.

Logistic Regression

Logistic regression predicts the probability of an outcome that can only have two values (i.e., a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g., predicting probabilities
- outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

In the logistic regression the constant (b0) moves the curve left and right and the slope (b1) defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio.

$$\frac{p}{1-p} = \exp\left(b_0 + b_1 x\right)$$

Finally, taking the natural log of both sides, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors. The coefficient (b_1) is the amount the logit (log-odds) changes with a one-unit change in x.

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

As mentioned before, logistic regression can handle any number of numerical and/or categorical variables.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

There are several analogies between linear regression and logistic regression. Just as ordinary least square regression is the method used to estimate coefficients for the best fit line in linear regression, logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target. After this initial function is estimated, the process is repeated until LL (Log Likelihood) does not change significantly.

$$\boldsymbol{\beta}^{\mathbf{1}} = \boldsymbol{\beta}^{\mathbf{0}} + [\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}]^{-\mathbf{1}}.\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{\mu})$$

$\boldsymbol{\beta}$ *is a vector of the logistic regression coefficients.*

$\boldsymbol{W}$ *is a square matrix of order N with elements* $n_i\pi_i(1-\pi_i)$ *on the diagonal and zeros everywhere else.*

$\boldsymbol{\mu}$ *is a vector of length N with elements* $\mu_i = n_i\pi_i$.

A pseudo R^2 value is also available to indicate the adequacy of the regression model. Likelihood ratio test is a test of the significance of the difference between the likelihood ratio for the baseline model minus the likelihood ratio for a reduced model. This difference is called "model chi-square ". Wald test is used to test the statistical significance of each coefficient (b) in the model (i.e., predictors contribution).

Pseudo R2

There are several measures intended to mimic the R2 analysis to evaluate the goodness-of-fit of logistic models, but they cannot be interpreted as one would interpret an R2 and different pseudo R2 can arrive at very different values.

Likelihood Ratio Test

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

The likelihood ratio test provides the means for comparing the likelihood of the data under one model (e.g., full model) against the likelihood of the data under another, more restricted model (e.g., intercept model).

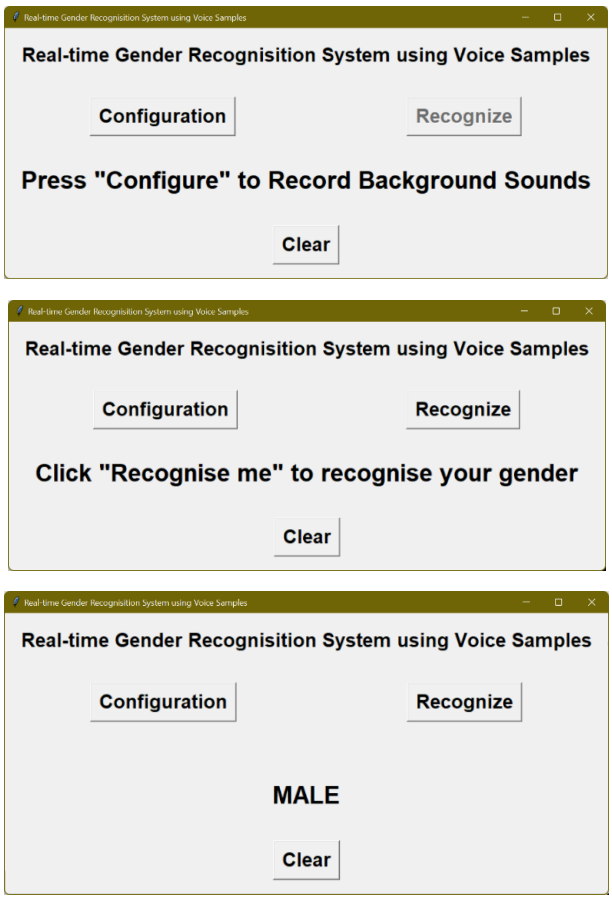
$$LL = \sum_{i=1}^{n} y_i ln(p_i) + (1 - y_i) ln(1 - p_i)$$

where *'p'* is the logistic model predicted probability. The next step is to calculate the difference between these two log-likelihoods.

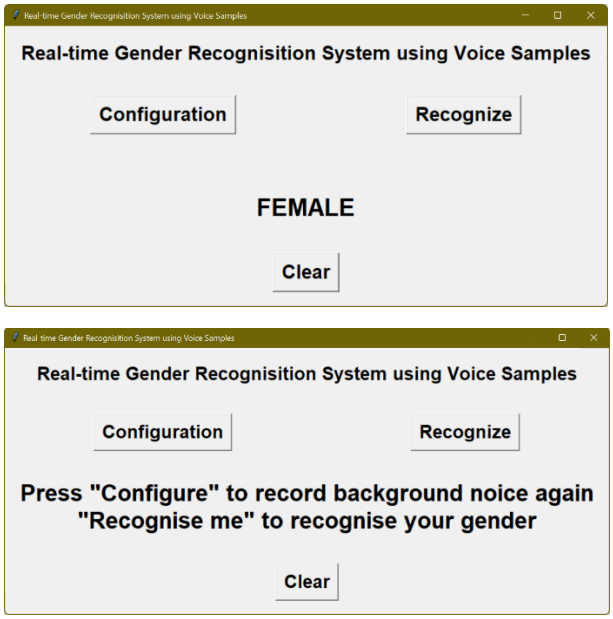
$$2(LL_1 - LL_2)$$

The difference between two likelihoods is multiplied by a factor of 2 in order to be assessed for statistical significance using standard significance levels (Chi^2 test). The degrees of freedom for the test will equal the difference in the number of parameters being estimated under the models (e.g., full and intercept).

4. Results



ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021



5. Conclusion

The model is found to be accurate for determining the gender using the voice and speech samples of the humans. The proposed model is less resource intensive and also has the ability to train faster with high accuracy due to use of SVM as the core. The results obtained by employing the combination of Principal component analysis and Support vector machine is close to the result obtained by using ML-deep learning. Also, the model can be trained by using smaller dataset unlike the case of the deep learning.

References

[1] Chen, X., 2022. Design of Political Online Teaching Based on Artificial Speech Recognition and Deep Learning. Computational Intelligence and Neuroscience, 2022.

[2] Jayne, C., Chang, V., Bailey, J., Xu, Q.A. (2022). Automatic Accent and Gender Recognition of Regional UK Speakers. In: Iliadis, L., Jayne, C., Tefas, A., Pimenidis, E. (eds) Engineering Applications of Neural Networks. EANN 2022. Communications in Computer and Information Science, vol 1600. Springer, Cham. https://doi.org/10.1007/978-3-031-08223-8_6

[3] Wani, T.M. et al. (2021). Multilanguage Speech-Based Gender Classification Using Time-Frequency Features and SVM Classifier. In: , et al. Advances in Robotics, Automation and Data Analytics. iCITES 2020. Advances in Intelligent Systems and Computing, vol 1350. Springer, Cham. https://doi.org/10.1007/978-3-030-70917-4_1

[4] Yadav, C.S., Yadav, M., Yadav, P.S.S., Kumar, R., Yadav, S., Yadav, K.S. (2021). Effect of Normalisation for Gender Identification. In: K V, S., Rao, K. (eds) Smart Sensors Measurements and Instrumentation. Lecture Notes in Electrical Engineering, vol 750. Springer, Singapore. https://doi.org/10.1007/978-981-16-0336-5_13

[5] Jha, T., Kavya, R., Christopher, J. et al. Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. Int J Speech Technol 25, 707–725 (2022). https://doi.org/10.1007/s10772-022-09985-6

[6] Altalbe, A. RETRACTED ARTICLE: Audio fingerprint analysis for speech processing using deep learning method. Int J Speech Technol 25, 575–581 (2022). https://doi.org/10.1007/s10772-021-09827-x

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

[7] Hourri, S., Kharroubi, J. A deep learning approach for speaker recognition. Int J Speech Technol 23, 123–131 (2020). https://doi.org/10.1007/s10772-019-09665-y

[8] Vaijayanthi, S., Arunnehru, J. (2021). Synthesis Approach for Emotion Recognition from Cepstral and Pitch Coefficients Using Machine Learning. In: Bindhu, V., Tavares, J.M.R.S., Boulogeorgos, AA.A., Vuppalapati, C. (eds) International Conference on Communication, Computing and Electronics Systems. Lecture Notes in Electrical Engineering, vol 733. Springer, Singapore. https://doi.org/10.1007/978-981-33-4909-4_39

[9] Xue, G., Liu, S., Gong, D. et al. ATP-DenseNet: a hybrid deep learning-based gender identification of handwriting. Neural Comput & Applic 33, 4611–4622 (2021). https://doi.org/10.1007/s00521-020-05237-3