Language Identification using Time-Frequency Image Textural Descriptors and GWO based Feature Selection by GPU Computation

Amit A. Chowdhury
Department of Electronics Engineering
Ramrao Adik Institute of Technology
Navi Mumbai,India
amitac2100@gmail.com

Rohit D. Phadke
Department of Electronics Engineering
Ramrao Adik Institute of Technology
Navi Mumbai,India
rohitphadke8@gmail.com

Balaji B. Dange
Department of Electronics Engineering
Ramrao Adik Institute of Technology
Navi Mumbai,India
dangebalaji143@gmail.com

***Abstract*—An ability to categorize and recognize a spoken language is an essential task in a multi-lingual society like India. Language identification (LID) is the process of identifying the language that is being spoken by some unknown speaker using a given speech sample. In this project, we have designed a hardware implementation of language identification using NVIDIA Tesla K80 (GPU) in real time. The frequency based features i.e., prosodic features are different for every language and also follows a certain patterns, which helps to recognize the language of an unknown speech. The proposed LID approach consists of four main stages. In the first stage, an audio sample is converted into a spectrogram visual representation which is a representation of the band of frequencies of a signal with respect to time. In the second stage, Rotational Invariant Complete Linear Binary Pattern (RICLBP) is used to extract the features from the spectrogram image. In the third stage, using Grey Wolf Optimizer (GWO), irrelevant and redundant features are removed and only optimal features are selected from the data set hence it helped to construct the classification model and the performance of the classifier is optimized. In the final stage, using Deep Neural Network, the pattern from the selected features are recognized by the trained classifier and classifies the unknown language sample to a known category.**

***Index Terms*—Language Identification, Spectrogram, RICLBP, Grey wolf optimization, Feature selection, Deep Neural Network.**

I. Introduction

Spoken language is a natural and powerful mode of communication and education. Spoken language identification is an important area in speech processing. In India, there are a total of 121 languages and 270 mother tongues across the country as specified under the Eighth Schedule to the Constitution of India. Overall, total 22 languages specified as major languages as per the Office of the Registrar General and Census Commissioner, India. According to Ethnologue [11], there are around 7,097 languages spoken around the world. Due to the development of information infrastructure and the Internet, there is an increase in the number of people connecting to a worldwide network. Dealing with businesses at a global level and communicating with people who use different languages to express themselves. Hence a language identifier is essential for better communication and useful in voice-based automated systems.

Language identification (LID) is the process of identifying the language that is being spoken by some unknown speaker using a given speech sample [12]. LID has a wide range of applications such as in natural language processing systems, information retrieval systems, document filtering systems, text mining applications, anywhere where you might need to work with more than one language seamlessly, identification of the language or encoding of web pages, multi-lingual spell checking, knowledge management systems, e-mail routing and filtering engines, content-based and language specific web crawlers and search engine. It is also used as a preprocessing step by the translation engine to determine the language of speech or text sample before translating it into other language.

II. Literature Survey

In a language identification task, various features of a speech like syntactic, acoustic, lexical, prosodic and phonotactic or any combination such features are used to differentiate a language. In [17], the first attempt made for identifying Indian languages. The algorithm uses Mel-frequency cepstral coefficients (MFCCs) for LID of four south Indian and one national (Hindi) language. In [18], the delay features, namely, normal group delay feature (NGD), auto-regressive group delay (ARGD), auto-regressive group delay with scale factor (ARGDSF) and one magnitude based feature (MFCCs) are used for LID. Features proposed in [18] improved the performance by combining all the group delay based systems separately with the MFCC based system. The auto-associative neural networks (AANN) for capturing language-specific features are explored in [19]. Prosodic features for obtaining the language-specific information are also presented in the study.

The language-specific prosodic features have also been explored in [5] for LID task. These prosodic features are extracted from syllable, word and sentence levels to capture language-specific prosodic knowledge. A two-level language identification system for Indian languages is developed using

MFCC features in [10]. Authors have modeled the system using hidden Markov model (HMM), Gaussian mixture model (GMM) and artificial neural networks (ANN). A method used Gaussian mixture models (GMMs) to model the language-specific excitation source information for LID task.

An approach based on linear predictive coding (LPC) and MFCC features for six Indian languages was proposed. To identify these languages, support vector machine (SVM) and random forest (RF) were used as classification techniques. A method is presented for 13 Indian languages using the MFCC and linear prediction cepstral coefficients (LPCC). For classification purpose, a deep neural network was used. PLDA based i-vector for a noisy database on making it robust and efficient is discussed. MFCC is employed to extract the features from three Indian languages, and multiclass-SVM was adopted for the classification.

III. Implementation of Indian language identification algorithm

The proposed language identification approach has been divided into two phases. In the first phase, the neural network is trained using spectrogram textural features generated from different known languages and in the second phase, the trained neural network is employed to determine how well it can classify an unknown language.

Figure 1 depicts the architecture of proposed Indian language identification algorithm using GWO feature selection and ANN classifier. From the Fig. 1, the training phase consists of four steps. Four Indian languages from IIIT-H Indic Speech Databases [13] are used for training and testing purposes. The input speech samples are in *wav* format. In the first stage, we convert these audio samples into spectrogram image. Spectrogram converts the signal into a time-frequency based representation. In the second stage, RICLBP is used to extract the features from the spectrogram image. RICLBP feature extraction step results in a feature matrix of dimension $(n \times k)$ for each language, where n is the total no. of samples and k is RICLBP feature vector dimension. In the third stage using grey wolf optimizer (GWO), optimal feature set is selected from the extracted features. It is a process of selecting a significant subset of features to create a more accurate model. And finally, these selected features are used to train the neural network classifier. Deep neural network as a classification model is employed to observe some hidden pattern in the feature matrices which will help us to recognize the languages of the unknown samples.

A. Spectrogram

A spectrogram is a visual representation of the band of frequencies of a signal with respect to time. In spectrogram, the energy content of a signal is expressed in terms of frequency and time. Frequency is represented on a vertical axis (Y-axis) and time on the horizontal axis (X-axis). The amount of energy in the signal at any given time and frequency is shown by the level of grey scale.

Depending on the size of the Fourier analysis window, different levels of frequency / time resolution are achieved resulting in (a) narrow band and (b) wide band spectrum. Spectrogram computation is briefly explained below.

- Divide the signal into sections of the same length. This section must be short enough so that the frequency of the signal does not change well within a segment.
- To get the short time fourier transform , window each segment and calculate its spectrum.
- Show the energy of each spectrum segment by segment in decibels and then show the magnitude intensity with the colormap.

In order to analyze the frequency content of a finite duration discrete time signal x with N samples, we use

$$x(k) = \sum_{N=0}^{N-1} x(n)e^{-}i\frac{2k\pi}{N}n \tag{1}$$

This can be interpreted as the Fourier Transform of the finite duration signal evaluated at the frequencies $f = k = N$. In order to obtain original signal x, inverse DFT is applied.

$$x(n) = \frac{1}{N}\sum_{N=0}^{N-1} x(k)e^{-}i\frac{2k\pi}{N}n \tag{2}$$

Figure 2 illustrates spectrogram of Hiindi, Marathi, Bengali and Malayalam languages. As it is evident from the figure that, for each language the spectrogram is different and shows the potential for language identification.

B. Rotational Invariant Complete Local Binary Pattern

Rotational Invariant Complete local binary pattern is generlised version local binary pattern (LBP) which is proposed by Z.Guo et al [11]. RICLBP is proved to be effective on texture analysis. In LBP, a local region represented by center pixel and difference between local. Whereas, in RICLBP a local region is represented by centre pixel and the difference between the values with local centre pixel with magnitude that is called as Local Difference Sign-Magnitude Transform (LDSMT). RICLBP has three different components, RICLBP-S indicates the sign (positive or negative) of difference between the centre pixel and local pixel, RICLBP-M indicates the magnitude of the difference between the centre pixel and local pixel and RICLBP-C indicates the difference between local pixel value and average central pixel value. RICLBP-S is nothing but normal LBP [13].

$$RICLBP - S_{P,R} = \sum_{p=0}^{p-1} s(g_p - g_c)2^p \tag{3}$$

where S(x) = $\begin{cases} 1 & \text{if} x \geq 0 \\ 0 & \text{if} x < 0 \end{cases}$

and g_c is the gray value of centre pixel and g_p is the value of is the value of its neighbours. RICLBP-M is calculated

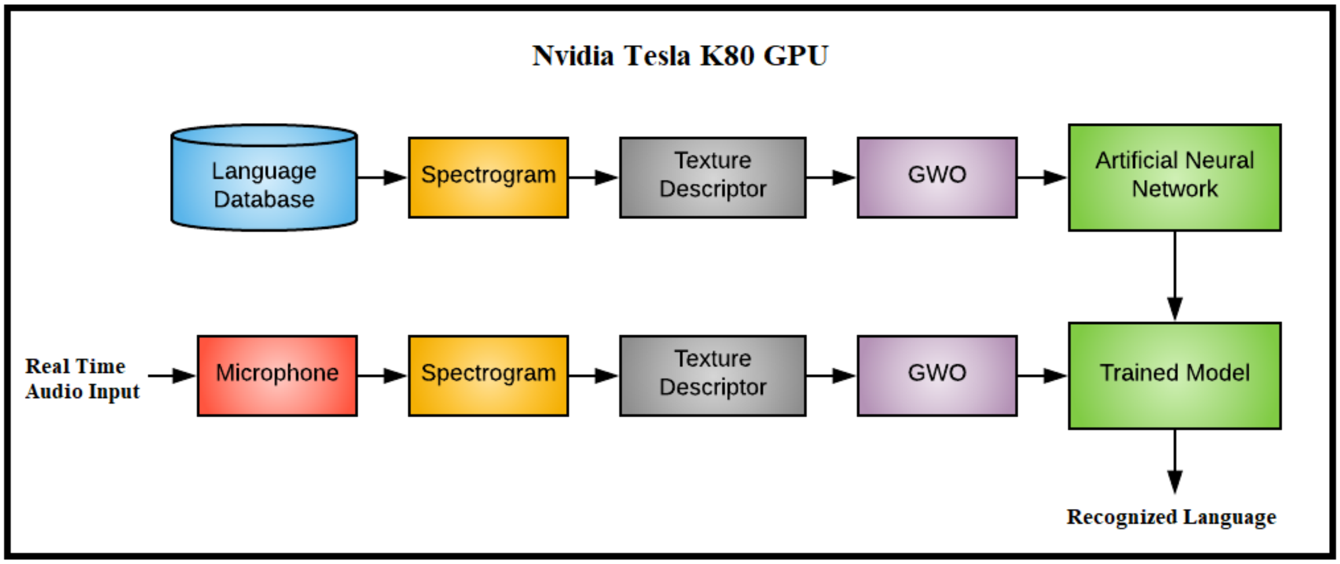

Fig. 1. Implementation of proposed Indian language identification algorithm

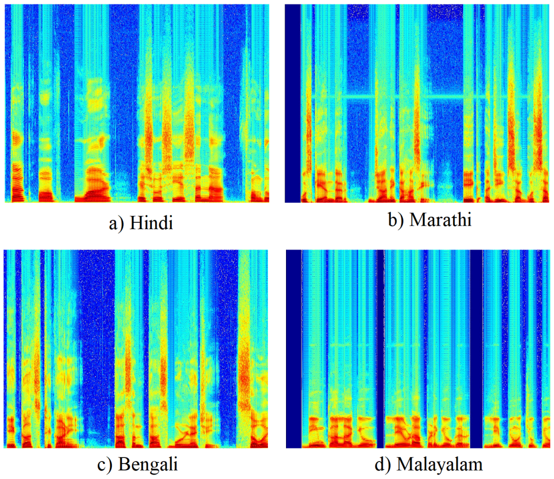
a) Hindi
b) Marathi
c) Bengali
d) Malayalam

Fig. 2. Spectrogram of different Indian languages

as same as RICLBP-S but it deals with the difference of the magnitude.

$$RICLBP - M_{p,r} = \sum_{p=0}^{p-1} t(m_p, c)2^p \tag{4}$$

where t(x,c) = $\begin{cases} 1 & \text{if } x \geq c \\ 0 & \text{if } x < c \end{cases}$ and c is a threshold to be determined adaptively and m_p is the magnitude component. The central image pixel also has discriminant information. Hence RICLBP-C is given by

$$RICLBP - C_{P,R} = t(g_c, c_i) \tag{5}$$

where, threshold c_i can be calculated as the average gray level of the whole image.

C. Feature Selection

The GWO the fittest solution is called the alpha (α) while the second and third best solutions are named beta (β) and delta (γ) respectively. The rest of the candidate solutions are assumed to be omega (ω). The hunting is guided by α, β, and γ and the ω follow these three candidates. In order for the pack to hunt a prey, they first encircle it. This encircling behavior can be modeled using the following equations:

$$\bar{X}(t+1) = \bar{X}_p(t) + \bar{A}\bar{D} \tag{6}$$

Where $\bar{D}$ is as defined in 2 and t is the number of iteration,$\bar{A}, \bar{c}$, are coefficient vectors $\bar{X}_p$ is the prey position and $\bar{X}$ is the gray wolf position.

$$\bar{D} = |\bar{C}\bar{X}_p(t) - \bar{X}(t)| \tag{7}$$

The $\bar{A}, \bar{C}$ vectors are calculated as in equation 3 and 4

$$\bar{A} = 2\bar{A}.\bar{r} - \bar{a} \tag{8}$$

$$\bar{C} = 2\bar{r}_2 \tag{9}$$

where components of $\bar{a}$ are linearly decreased from 2 to 0 over the course of iteration and r_1, r_2 are random vectors in [0,1]. The hunt is usually guided by the alpha.The beta and delta might also participate in hunting occasionally. In order to mathematically simulate the hunting behavior of grey wolves,the alpha(best candidate solution)beta,and delta are assumed to have better knowledge about the potential location of prey the first three best solutions obtained so far and oblige the other search agents (including the omegas) to update their position according to the position of the best search agents.

$$\bar{D}_\alpha = |\bar{C}_1\bar{X}_\alpha - \bar{X}|, \bar{D}_\beta = |\bar{C}_2\bar{X}_\beta - \bar{X}|, \bar{D}_\delta = |\bar{C}_3\bar{X}_\delta - \bar{X}| \tag{10}$$

TABLE I
EXPERIMENTAL RESULTS

Input language	Output language				Accuracy (%)
	Hindi	Marathi	Bengali	Malayalam	
Hindi (302)	299	0	2	1	99.01
Marathi (289)	0	289	0	0	100.00
Bengali (293)	1	0	292	0	99.70
Malayalam (272)	0	0	0	272	100.00
		Overall accuracy			99.68

$$\bar{X}_1 = |\bar{X}_\alpha - \bar{A}_1\bar{D}_\alpha|, \bar{X}_2 = |\bar{X}_\phi - \bar{A}_2\bar{D}_\phi|, \bar{X}_3 = |\bar{X}_\delta - \bar{A}_1\bar{D}_\delta| \quad (11)$$

$$\bar{X}(t+1) = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3} \quad (12)$$

A final note about the GWO is the updating of the parameter $\bar{a}$ that control trade off between exploitation and exploration. The parameter a is linearly updated in each iteration to range from 2 to 0 according to the equation .

$$\bar{a} = 2 - t.\frac{2}{Max_{iter}} \quad (13)$$

where t is the iteration number and Max_{iter} is the total number of iteration allowed for the optimization.

IV. EXPERIMENTAL RESULTS

Using GWO with the combined extracted feature, the dimensional size of the feature vector is reduced from 136 to 82, i.e., around 40% of the irrelevant and redundant attributes are removed. The confusion matrix computed using GWO feature selection. Comparing the results obtained without employing feature selection, the accuracy of four languages has marginally improved which leads to an increase in the overall accuracy of the system. The highest accuracy achieved for a language is Marathi and Malayalam whereas lowest is for Hindi. Using GWO an overall accuracy of 99.68% is achieved improving the overall accuracy rate at the same time keeping the feature vector dimensionality low.

V. CONCLUSION

We have demonstrated the image textural descriptors extracted from spectrogram image for Indian language identification and the use of GWO feature selection. The frequency band for every language is different from each other i.e, prosodic features and it follows certain distinctive textural patterns for each and every language, which helps to recognize the language of an unknown speech. These patterns were successfully exploited by the GWO by selecting only the optimal features which helped to train the deep neural network efficiently. Hence the GWO approach proves to be robust, reliable and fast compared to other approaches performed in this study.

REFERENCES

[1] Vicky Kumar Verma and Nitin Khanna, ”Indian Language Identification Using K-Means Clustering and Support Vector Machine (SVM)”.

[2] Chithra Madhu, Anu George and Leena Mary, ”Automatic Language Identication for Seven Indian Languages using Higher Level Features”.

[3] Sreedhar Potla and Vishnu Vardhan. B, ”Spoken Language Identification using Gaussian Mixture Model-Universal Background Model in Indian Context”.

[4] V.Ramasubramanian, A.K.V.S. Jayram,T.V. Sreenivas, ”Language Identification using parallel sub-word recognition an ergodic HMM equivalence”(Geneva, Switzerland) 2003.

[5] L. Mary, Multilevel implicit features for language and speaker recognition, Ph.D. dissertation, Indian Institute of Technology Madras, India (2006).

[6] Nora Barroso, Karmele Lpez de Ipia, Carmen Hernndez, Aitzol Ezeiza, Manuel Graa, ”Semantic speech recognition in the Basque context Part II: language identication for under-resourced languages”.

[7] Rong Tong, ”A Target-Oriented Phonotactic Front-End for Spoken Language Recognition”.

[8] Sabato Marco Siniscalchi, Jeremy Reed, T. Svendsen, Chin-Hui Lee, ”Universal attribute characterizaton of spoken languages for automatic spoken language recognition”.

[9] Timothy J. Hazen and Victor W. Zue, ”Segment-based automatic language identification”.

[10] S. Jothilakshmi, Sengottayan Palanivel and V. Ramalingam, ”A hierarchical language identification system for Indian languages”, Digital Signal Processing, Vol.22, No.3, pp.544–553, 2012.

[11] https://www.ethnologue.com Accessed on 3 August, 2018.

[12] Martine Adda-Decker, Automatic Language Identification,2008.

[13] http://festvox.org/databases/iiitvoices/ Accessed on 30 December, 2018.

[14] Zhenhua Guo, Lei Zhang and David Zhang, “A Completed Modeling of Local Binary Pattern Operator for Texture Classification”, IEEE Transactions on Image Processing, Vol.19, No. 6, pp. 1657–1663, 2010.

[15] Seyedali Mirjalili, Seyed Mohammad Mirjalili,and Andrew Lewis, “Grey Wolf Optimizer”.

[16] Bakshi Aarti and Sunil Kumar Kopparapu, ”Spoken Indian Language Identification: a review of features and databases”.

[17] Jyotsana Bal leda, Hema A Murthy, and T.Nagarajan, “Language identification from short segments of speech”, Sixth Int. Conf. on Spoken Language Processing, Vol.3, pp.1033-1036, 2000.

[18] Arup Kumar Dutta and K. Sreenivasa Rao, “Language identification using phase information”, International Journal of Speech Technology, pp.1–11, 2017.

[19] L.Mary, K. S. Rao, S. V. Gangashetty, and B. Yegnanarayana, “Neural Network models for capturing duration and intonation knowledge for language and speaker identification”, 8th Int. Conf. on Cognitive and Neural Systems, pp.1-4, 2004.

[20] D.-S. Huang and H.-J. Yu, Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 2, pp.457467, 2013.

[21] Nagaraja S., Prabhakar C.J. and Praveen Kumar P.U., Complete Local Binary Pattern for Representation of Facial Expression Based on Curvelet Transform, 2013.