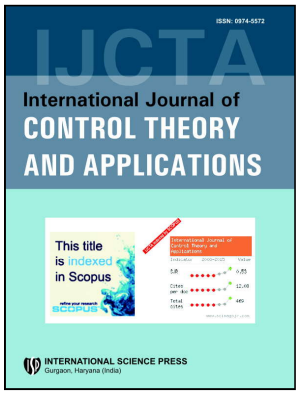

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 18 • 2017

Performance evaluation through the use of different kernel functions in SVM for the identification of native Indian Languages

Chandrakanta Mohapatra[1], Sujata Dash[2], Ansuman Mishra[3] and Umakanta Majhi[4]

[1] *PhD Scholar, Dept.of Computer Application, North Orissa University, Orissa, India, Email: ck.mohapatra@gmail.com*
[2] *Reader, Dept.of Computer Application, North Orissa University, Orissa, India, Email: sujata238dash@gmail.com*
[3] *Sr.Lecturer, Dept. of CSE, BIET, Bhadrak, Orissa, India, Email: ansuman.me@gmail.com*
[4] *Assistant Proffesor, Dept. of CSE, NIT, Silchar, Assam, India, Email: umakantmajhi@gmail.com*

Abstract: As automatic language identification plays a vital role in multi-lingual platforms, and speech becomes a common user interface so, there has been a wide research in this field to enhance the performance and accuracy of identification, but still many challenges are there to address. In the way of improvement in accuracy, the Support Vector Machine has been used mostly for the discrimination purposes. There different types of kernel functions are available to map the data in input space into a higher dimensional feature space. Here in this paper, different kernel functions of support vector machine are applied to build models for identifying native Indian languages such as Odia, Hindi, Bengali and Telugu from speech. The accuracy of the models is evaluated with a speech dataset consisting of four native Indian languages. It is observed from the experiment that performance of RBF model is promising in comparison to other two models. Besides this, the result of the study is compared with the findings of other authors to establish its efficiency.

Keywords: LID-language Identification, SVM-Support vector Machine, GMM- Gaussian Mixture model, MFCC-Mel frequency cepstral coefficients, LPCC-Linear Perceptual Cepstral coefficients , RBF-Radial basis function.

1. INTRODUCTION

Language identification (LID) is the process of identifying the identity of language corresponding to a given spoken utterance by a system. The main aim of automatic Language Identification (LID) is to minutely and accurately identify the language being spoken. Language identification has numerous applications in a wide range of multi-lingual services. An example is the language identification system used in Automatic Voice controlled travel information retrieval system [6]. However, there are many benefits to be gained from making LID an automatic process, some of the benefits include the reduced costs and shorter training periods associated with the automated system. For multiple human language identification services, several people would need to

be trained to properly recognize a set of languages whereas the LID system can be trained once and then run on multiple machines simultaneously [7]. For the Identification of language ,it required to extract the speech related features and through the speech feature the modeling has to be done, and later these models can be used for the verifying the test utterance to know the accuracy in identification. So in the first phase different speech related features need to be generated, there different parameterization techniques such as MFCC,LPCC,PLP can be used for that purpose but Commonly the spectral features such as MFCCs are used as the feature vector of the speech signal, there we can also use LPCC features but in most cases MFCC features provide better performance [2][10]. In this paper, MFCC is used as the feature extraction method. It is also possible to use HMM (Hidden Markov Model) or GMM (Gaussian Mixture Model) modeling through the MFCCs, but in case of GMM it needs more data of speech utterances, which cost more in the process of model building and further verification [9]. But SVM provides an elegant way to classify data and considered as a best classifier when small amount of training data or speech features are available. In addition to this, it has the capability to deal with non-linear problems employing kernel functions [3][6].

The paper is organized as follows: section 2 explains about the feature extraction method followed by section 3 which explains about language modeling using different kernels of SVM. Then section 4 discusses about experimental setup and speech datasets where as results of the experiment are discussed in section 5. Section 6 concludes the paper followed by future projection in section 7.

2. FEATURE EXTRACTION PROCESS

For any classification task, it need to generate features containing the characteristics of the raw speech. So Feature extraction is the process of extracting speech features like amplitude, frequency components, prosodic, phonotic, acoustic features, that represents the utterance being spoken. An acoustic speech feature vector is the compact representation of the raw speech which can be formed through MFCC, and again MFCC can be used for the spectral analysis of the speech signal[16]. As the frequency component of speech signal is non-linear, therefore each tone with actual frequency f is measured in Hz and pitch interms of Melody scale [5]. The Mel scale can be calculated as

$$f_{mel} = 2595 \times \log(1+f/700) \quad (1)$$

MFCC is mainly focused on converting linear spectrum into non-linear. Mel scale filter has series of triangular uniform overlapping filters with constant bandwidth of 100 and centre frequency at 50, like the perceived human auditory system [8][15]. The spacing as well as bandwidth of the particular filters is determined by a constant mel-frequency interval. This is accomplished by computing mel-cepstral coefficients, and can be obtained by applying inverse DFT to the log energy output of the filter bank [10][12]. The series of phases required for the conversion is depicted through the figure given below.

The stages of the MFCC process are discussed underneath:

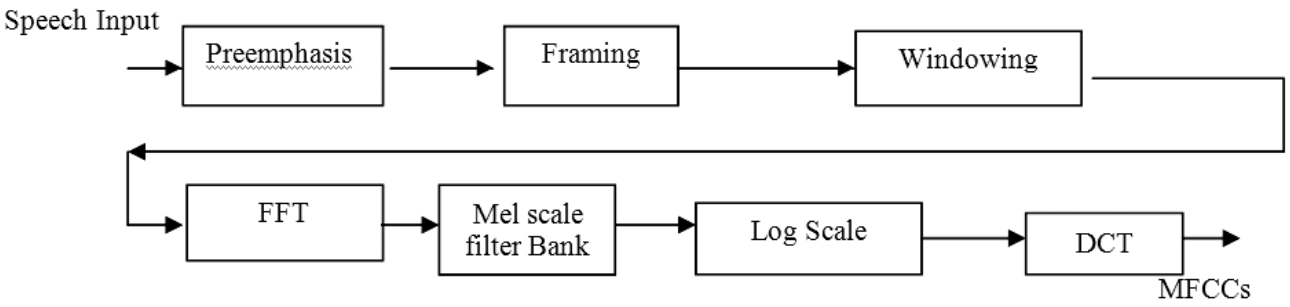

Figure 1: MFCC parameterization process

- Preemphasis-This is the first stage of MFCC to boost the energy in high frequency, as there are more energy in lower frequencies than higher frequency. So it causes spectral tilt, to avoid that preemphasis is required thereby it can make more information available from these higher formant to the model and resulted in increased accuracy. This can be performed by using filters
- Framing-The process of forming frames by overlapping, so that there will not be any discontinuities between the frame, the width is generally 30ms and an overlap of 20ms, i.e a shifting of 10ms [4].
- Windowing-As the speech signals are non-stationary, so the statistical spectral properties are not constant across time, so we have take a small window for feature extraction, with an assumption to make it stationary, to find the characteristics and characterize the subphones, we usually use hamming window which shrinks the values towards zero at window boundary, and avoid the discontinuities resulted in case of rectangular window. If the hamming window defined as w(n), 0d"nd"N-1,where N is the number of samples ,y(n) is the output, x(n) is input signal ,then the resultant signal after windowing operation is y(n)=x(n)×w(n) [5][15].
- FFT-This is required to extract the spectral information for the windowed signal and to know the energy that the signal contain at different frequency band. Through this we can extract spectral information for discrete frequency band, and for discrete time. Discrete fourier transform is used for this purpose, commonly there used FFT to calculate DFT, which convert each frame of N sample from time domain into frequency domain [4].
- Mel scale filter bank-Here filters are non-uniformly spaced on the frequency scale, with more filters in low frequency region and fewer filters in high frequency region; the bank of filter is applied to the spectrum as per mel scale.
- Log scale-Through this logarithm value of the square magnitude of the output of mel filter bank is calculated, as the log compress dynamic range of values.
- DCT-This is the process of converting the log mel spectrum into time domain, which resulted in mel frequency cepstral coefficients, also called acoustic vectors .it require fourier analysis.

3. LANGUAGE MODELING THROUGH SVM

SVM is a two class classifier used for the classification task, multiple pattern classification problem can be solved through this. SVM provides a method that makes decision about constructing a hyperplane that separate two class optimally, In case of linear classifier it separates a set of objects into respective groups with line, but most of all are complex and need optimal separation i.e. correctly classify new object on the basis of the example that are available [8]. In case of linear classification as depicts in the fig.2 ,which present an optimal hyperplane for linear classification , triangle and square symbols represent the sample of two class, H represents the hyperplane , the distance between the closest data point and the separating hyperplane is called the margin of separation, the distance between A and B called the margin, there may be many hyperplane that separate the samples but only one of them attain the maximum margin i.e. an optimal hyperplane tries to attain largest margin between classes,as the goal of SVM is to find a particular hyperplane that maximizes the margin of separation [5][13]. For linear separable data $\{x_i, y_i\}$, $x_i \in R^d$, $y_i \in \{-1, 1\}$, here x_i, y_i are linearly separable training sample. The optimal hyperplane is calculated according to the formula i.e.

$$F(x) = w.x + b=0 \quad (2)$$

where w is the weight vector. The optimal hyperplane is calculated according to the maximum margin [5]. The solution for the optimal hyperplane w_0 is linear component of the small subset of data x_s, $x \in \{1..N\}$ known as the support vectors. The full separation requires a curve classification task based on drawing separating lines to distinguish between object of different class membership are known as hyperplane classifiers. For non-linear

separable data, no hyperplane exists ,so as to satisfy the inequality .There is a hyper parameter used to specify the effect of minimizing the empirical risk and maximizing the margin, the formulation can be formed as

$$\text{Max } (\sum\alpha_i+\sum\alpha_i\alpha_j\, y_iy_j,x_i,x_j) \quad (3)$$
$$\text{Subject to } 0\leq\alpha_i\leq c\ \sum\alpha_iy_i=0$$

where α is the Lagrange multiplier of the ith constraint for the optimized problem. The optimal plane can be formulate as

$$W_0=\sum i\ \alpha_iy_ix_i \quad (4)$$

The training data for classifier are not always linearly separable, to handle such condition, kernel is being introduced, through this SVM can perform a non-linear mapping from lower dimensional to higher dimensional space [8].so that the training sample which are not linearly separable can be linearly separable in higher dimension feature space[1].

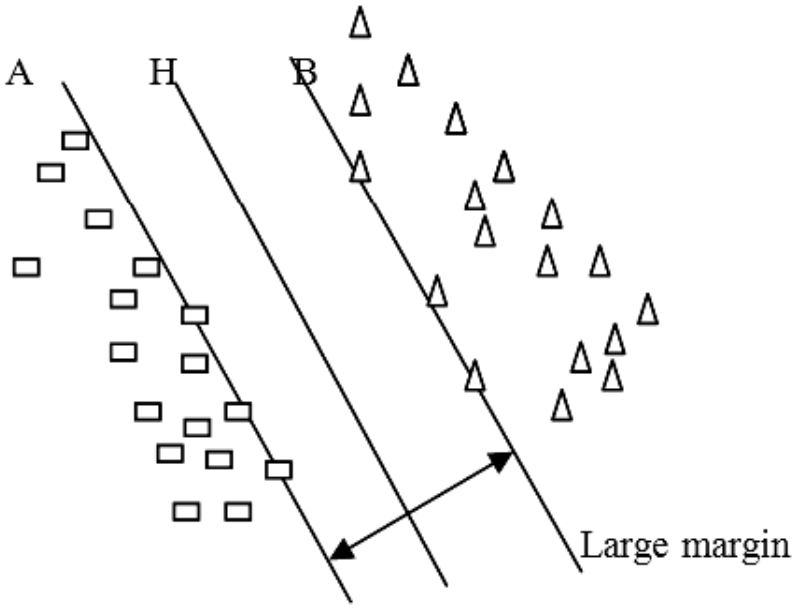

Figure 2: Optimal hyperplane for linear classification

According to mercer's theory there exist a kernel function which satisfies

$$K\ (x_i,x)=\varphi(x).\varphi(x_i) \quad (5)$$

Where $\varphi(x)$ represent the image induced in feature space due to the input vector [1]. The kernel functions are described below

A. Linear function-This function is formulated by the inner product of two vectors in low dimension space.

$$\text{Kernel}(x_i,x_j)=x_i.x_j \quad (6)$$

B. Polynomial function – This function is suitable for normalized training sample

$$\text{Kernel}\ (x_i,x_j)=(x_i.x_j+1)^p \quad (7)$$

where p is the degree of polynomial

C. Radial basis Function- This is in Gaussian form. The formulation is given by

$$\text{Kernel}(x_i, xj) = \exp[-1/2(||xi-xj||/\sigma)^2] \quad (8)$$

where σ is the width of the radial basis function.

4. EXPERIMENTAL SETUP

The experiment is performed using a speech database which was created by collecting recorded speech from broadcasting television channels using DISHTV direct to home connection. The speech corpus was created mainly from the news bulletins and discussion. The database is consisting of four native Indian languages i.e Hindi, Oriya, Bengali and Telugu. For each language, speech is collected from a different speaker. So for each of the language in the group, around 35 to 40 minutes of speech data is collected. We have used the Audacity tool for recording the speech signal using the TV tuner card through the computer. The main reason for choosing television broadcast speech content is, in case of recording of speech data from various persons in different languages is a time consuming process and the recorded signal may contain noises from the device and environmental interference, which may result as noisy speech signal and may tends in degradation of performance. Therefore, to obtain clear and noise free speech, television broadcast speech was taken as its speakers are professional and matured. Again, the recorded speech may contain some music or overlapping voice, so it needs to be edited to remove such things from speech before it gets processed. The speech signal is recorded in sampling rate of 16 KHz .The feature extraction is performed through Matlab R2013a and SVMTorch is used for language model building and classification process. The experiment is conducted inorder to evaluate the performance of different kernel functions of SVM. Recorded speech from different speakers for four different native Indian languages such as, Hindi, Odia, Bengali and Telugu are taken and MFCC is applied to extract feature vectors from the speech. The language models are built on the training data which is 70% of the whole dataset. The remaining 30% is used for the validation purpose.

5. RESULT & DISCUSSION

In this experiment, melcepst function is used to extract the cepstral co-efficient. By using these cepstral co-efficient i.e. the speech spectral vectors, different models for each languages are developed through the SVM Tool. The input speech is used to train the models. The number of feature vectors used in training and testing purposes are presented in the TABLE I. While building the model we have also changed the parameter for kernels i.e –t <int > ,where <int> represents the value for kernel type '0' for the linear,'1' for polynomial and '2' for Gaussian kernel or RBF kernel,along with this we can use –d <int> for the degree of polynomial in polynomial kernel function and –st <float> for the ó of the RBF kernel function. After building the model by varying the parameters, we have combined them together to form language identification system, through which the discrimination of language is to be carried out. We can verify the language model by the test utterance kept earlier, for that again we have to convert the speech contents for test utterance into cepstral features and then have to match it onto the models build earlier to get the match in classification .we can calculate the accuracy

Table 1
List of feature vectors

List of feature vectors	*Number of feature vectors*		
	Language	*Training*	*Testing*
1	Hindi	14937	1782
2	Oriya	11592	1681
3	Bengali	16040	2054
4	Telugu	17095	2141

percentage by the number of classified features over the total feature vectors presented for training. So for each cases i.e linear, polynomial and RBF different set of language models were build and in the same way the test utterance is tested over the language models, for calculating the accuracy percentage over the data set taken.

Here SVMs trained with polynomial kernel by taking the polynomial degree p=10,and the SVMs trained by RBF kernel through the ó value as 0.1.We have tested the models by using test speech data of 4 different languages.The percentage of accuracy noted from the discrimination models are presented in TABLE II.

Table 2
Accuracy resulted from models

	Accuracy % on different Kernel function used		
Language models	*Linear*	*Polynomial*	*RBF*
Hindi	64.36 %	72.56 %	83.45%
Oriya	61.23 %	70.89%	82.38%
Bengali	52.91%	63.73%	74.29%
Telugu	57.11 %	66.82%	78.13%

If we compare the result with the observed result of D.Ben et.al[1], we observed that in our experiment we got higher performance in RBF kernel and got less percentage in linear kernel as compared to the experiment carried out by Aditya Bhargav et.al[11] .But as observed by Joachim[14], we also observed from the result ,that it shows considerable better performance in polynomial and RBF kernel. so we can draw an analogous with that research findings.In our experiment we had different kernels, the kernel that performs best is the radial basis function (RBF).Again if we compare the result with earlier research paper [11] where the authors had used the Transfermarkt corpus consisting of European soccer player names with 13 possible languages, with separate list for last names and full names. They had used character level language modeling. The result generated from the SVM language model through different kernel functions are listed below in TABLEIII. Their classification was based on the last names and then through the full names as the testing sample for recognition.

Table 3
LID Accuracy on transfermarkt corpus

	Accuracy in percentage	
Language model methods	*Based on Last name*	*Based on full name*
Linear SVM	56.4 %	79.9%
RBF SVM	55.7%	78.9%
Sigmoidal SVM	56.2%	78.7%

By comparing the above findings with our result, we can summarize that the performance in identification can be improved through the speech corpus.

6. CONCLUSION

This paper covers the utilization of different aspect of speech information, for identifying the language. We obtained the feature vectors for different languages, that used to train the system in order to build models for different languages by using SVM .we had also varied different parameters to get the different models based on the different kernel types. We have attempted to find best kernel type. So to investigate that, we have test the models over the test speech to verify the accuracy percentage, and found that RBF kernel function resulted in better performance as compared to other kernel functions i.e linear and polynomial.

7. FUTURE WORK

Here all language models have used the language recorded from a single speaker, so the model consisting of less linguistic features, rather than more information about speaker. so in order to get more linguistic features, we can add different speakers in each language group and can use k-cross validation ,where k-1 sub sample can be used for training and a single sub sample can be used for validation. Accuracy of the model can be measured through mean of performance, again we can also select optimized parameters for every kernel function by varying the adjustable parameters to achieve higher accuracy for the model.

REFERENCES

[1] D.Ben, Ayedmezghani, S.Zribi Boujelbere, N.Ellouze "Evaluation of SVM kernel and conventional machine learning Algorithms for speaker identification"International journal of hybrid information technology,vol 3,July 2010.

[2] Bo Yin1, Eliathamby Ambikairajah1, Fang Chen2 "Combining Cepstral and Prosodic Features in Language Identification" School of Electrical Engineering and Telecommunications UNSW1, National ICT Australia Ltd. The 18th International Conference on Pattern Recognition (ICPR'06) © 2006 IEEE.

[3] Yan Deng, Jia Liu " Automatic Language Identification using support vector machine & Phonetic N-gram" Tsinghua National Laboratory for Information Science and TechnologyDepartment of Electronic Engineering, Tsinghua University, Beijing 100084, ChinaE-mail: y-deng05@mails.tsinghua.edu.cn ©2008IEEE .

[4] W. M Chambell,E.singer ,p. A. tores charasequela,D.A Reynolds"Language recognition with support vector machine"MIT Lincoln laboratory,Lexington,M A USA.By DOD Air Force US 2009

[5] Bo yu,Haifeng Li,Chunying Fang "Speech emotion Recognition based on optimized support vector Machine" Journal of software Vol.7,No.12 December 2012.

[6] Eliathamby Ambikairajah,Haizhou Li, Liang Wang,Bo Yin, and Vidhyasaharan Sethu "Language Identification a tutorial" IEEE CIRCUITS AND SYSTEMS MAGAZINE SECOND QUARTER 2011.

[7] Marc A. Zissman, "Automatic Language Identification of Telephone Speech" ,THELINCOLN LABORATORY JOURNAL VOLUME 8. NUMBER 2. 1995.

[8] A.Miton, S. shrmy Roy,S.Tamil Selvi "SVM scheme for speech Emotion Recognition using MFCC feature"International journal of computer Application(0975-8887)volume 69-no.9,May 2013.

[9] Khe Chai Sim, Haizhou Li," On Acoustic Diversification Front-End for Spoken Language Identification" IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 16, NO. 5, JULY 2008.

[10] Namrata Dave "Feature extraction methods LPC,PLP & MFCC in speech Recogntion"GHPCE,Gujurat Technological University,IJANET, july 2013.

[11] Aditya Bhargava ,Grzegorz kondrak"Language Identification of names with SVM",University of Alberta Canada, Association for Computational Linguistics, USA ©2010.

[12] Vicky kumar verma,Nitin khanna. "Indian language identification using k-mean clusetering and support vector machine"Graphics era University, Dehradun©2013 IEEE

[13] A.B.M.S Ali,A Abraham,"An empirical comparison of kernel selection for support vector machine"soft computing systems,IOS Press publisher,Amsterdam 2002,pp321-330.

[14] T.Jaochim,"Text categorization with support vector machine:learning with many relevant features"10^{h} Europian conference of machine Learning ECML-98,pp 137-142

[15] Induja K,Nibesh k,P C Raghuraj "Text Based language identification using SVM"International Journal of computational Linguistic and natural language processing,Issue 4-9-2014

[16] Abhijeet Sangwan, Mahnoosh Mehrabani, John H. L. Hansen " Language Identification Using A Combined Articulatory Prosody Framework" University of Texas at Dallas, Richardson, Texas, U.S.A.,supported in part by USAF©2011 IEEE.