An Integrated Spoken Language Recognition System Using Support Vector Machines

Garima Vyas, Malay Kishore Dutta
Dept.of Electronics and Communication Engineering
Amity University, Uttar Pradesh, Noida -201303, India
gvyas@amity.edu, mkdutta@amity.edu

Abstract: **An automatic Language Identification (LID) is a system designed to recognize a language from a given spoken utterance. The spoken utterances are classified according to the pre-defined set of languages. In this paper a LID system is designed for two different languages namely English and French. The classification of an audio sample is done by extracting Mel frequency cepstral coefficients (MFCCs) and putting them on support vector machines with radial basis function kernel. The proposed framework is speaker-independent. This scheme was tested on a database of multi-lingual speech samples. The language identification accuracy is found to be 92% for French and 88% for English.**

Keywords — *Language identification; Support Vector Machine; Radial basis function; Mel Frequency Cepstral Coefficients; Audio Clips.*

I. INTRODUCTION

Speech is one of the most natural and efficient means of communicating information because of which the researchers have made great efforts to extract useful features from a speech utterance. Automatic Language Identification is a process of categorizing a speech sample as corresponding to one of the previously encountered languages. The process of Language Identification [1] is independent of context, content, task of vocabulary and robust with regard to speaker identity [7] , sex, age and also to noise and any distortion introduced in or by the communication channel [3]. Language identification plays a very critical role in various speech related applications. The main challenge in this research is that speech of language is given as input and the system has to identify the language. If the person has familiarity with a given language it becomes easier to identify a language from its short utterance compared to machines. The characteristics that make one language differ from another language are phonology, morphology, syntax and prosody. LID [7], [8], [10] has various applications where one application could be a telephone based front-end system whose main aim is to route the call to the appropriate caller who is fluent in that language. Other applications of language identification system would be in speech-to-speech translation, shopping, airports and other commercial areas.

The design of a LID [1] system has two main components: Extraction of interested features from a speech utterance and their classification. Audio Features can belong to many domains like temporal features, phase domain, cepstral [4] features etc. For LID approach, cepstral features are commonly used. Hence, MFCCs [3], [4] are extracted and used for classification. The classification based on support vector machine (SVM) can be divided in two phase: training and testing .SVMs are used as they are promising nonlinear and nonparametric classification technique which has shown good results in various fields like medical diagnostics, optical character recognition, etc. Overall, SVMs are spontaneous; well-founded, and is practically successful, reliable and efficient.

The major contribution of the proposed technique is to recognize the language spoken by a multi-lingual speaker. The design of LID is speaker independent and is robust to noise. The design of the system is also gender independent. Hence, it can find its application in language translation, indexing of online clips on the basis of language(s) etc. This work has huge scope to be extended to speech-to speech translation.

The rest of the paper is organized as follows. Section II describes the methodology followed. Section III extends the feature extraction method. Training and testing by SVMs are described in section IV. Section V shows the experimental set up and results. Finally the next section draws the conclusion and future scope of the paper.

II. METHODOLOGY

The database of 50 sentences is created for English and French language each. In each database, 25 audio clips have male voice and 25 clips are recorded with female voice. 10 sentences of each language are used for training and rest 40 is used for testing. To achieve the above mentioned objectives the methodology adopted is explained by the following flowchart shown in Fig 1. The input audio clip is first pre-processed. In pre-processing the raw sample file with duration of 15secs is clipped. After this, the clipped audio is converted into .wav format. Then, MFCCs [3], [4] were extracted from the formatted audio for both English and French. Then classification of the language is done using SVM [6], [9] with radial basis function (RBF) kernel as pattern recognition classifier. The output of the SVM classifier denotes the correct language.

978-1-4799-5173-4/14/$31.00 ©2014 IEEE

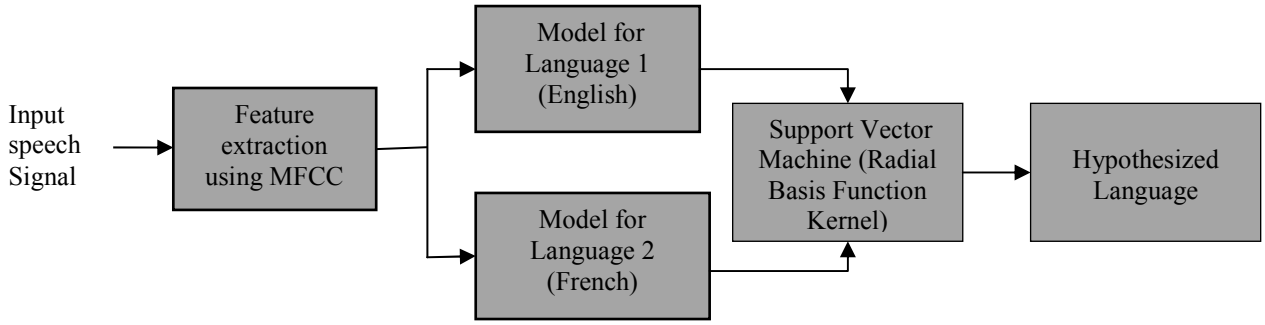

Fig 1: Proposed method for LID

III. FEATURE EXTRACTION USING MFCC

Only voiced segments of the speech signal are processed for MFCCs extraction. The procedure to determine MFCCs [3] is described as follows:

The input audio signal length is chosen to be 10 seconds and the signals are segmented in 20~ 40 ms frames with an optional overlap of 10~20 ms. Framing is done to make the audio signal statistically stationary. Now the MFCC is calculated in a sequential order as per the following steps:

Step 1: To minimize the distortion in the spectra and to reduce the signal's discontinuity, the framed signal is passed through the window. Hamming window is applied, to get framed signal, described in (1) is:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (1)$$

Step 2: On each windowed signal, the Discrete Fourier Transform (DFT) is performed to obtain the magnitude frequency response. The equation for DFT is:

$$X(k) = \sum_{n=0}^{N-1} X(n) e^{-\frac{j2\Pi kn}{N}}, \quad 0 \le k \le N-1 \qquad (2)$$

Step 3: Now, the windowed frequencies are passed through the series of triangular bank bass filters. These filters smoothes spectrum and also reduces the size of the feature points. These filters are equally spaced along the mel scale. Humans are much better at perceiving small changes in the pitch at low frequencies than at higher frequencies. Instilling this scale makes our features match more closely with what humans listen.. Hence, the conversion of frequency to mel scale is must and is given by (3) as follows:

$$M(f) = 1125 \times \ln\left(1 + \frac{f}{700}\right) \qquad (3)$$

Step 4: In the last step the output filtered frequencies are again converted back into a time domain by applying discrete cosine transform (DCT) on it. A DCT is chosen because it has excellent energy compaction. Eq. (4) describing DCT is:

$$Y(k) = \sum_{n=0}^{N-1} \cos[\frac{\Pi}{n}(n + \frac{1}{2})k] \quad 0 \le k \le N-1 \qquad (4)$$

For efficient performance of the system, we add the logarithmic energy. Taking logarithmic of energy compresses the dynamic range and makes the feature less variable to acoustic coupling changes. The result of the conversions is called Mel Frequency Cepstrum Coefficients.

IV. CLASSIFICATION USING SVM

SVM is a machine learning technique which is supervised in nature. SVM's [9] is a new technique used mainly for binary classification. The basic SVM [11] contains no prior knowledge of the problem. It classifies the data points into two classes -1 and +1. The main objective behind SVMs is to find a hyperplane which separates a d-dimensional data perfectly into two subgroups. The classification with the maximum margin is the best. Fig.2 shows model for SVM classifier.

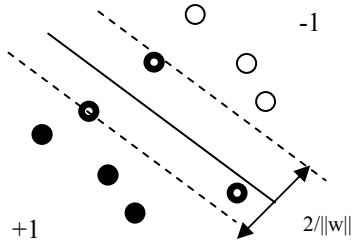

Fig 2. Classification using SVMs

The data points lying on the margin are known as support vectors. If these support vectors are deleted the width of the margin will get increased and hence better classification can be done. All the points on one side of the hyperplane are marked with class +1 and all data points on the other side are marked as -1.

For the testing phase, the decision rule for linear SVM is given by (5):

$$\overline{w} \cdot \bar{u} + b \geq 0 \qquad (5)$$

Where $\bar{w}$ the perpendicular to the median line of the street is, $\bar{u}$ is the unit vector and b is the constant. If the above equation satisfies then the data point belongs to +1 class. For some positive and negative sample the above equation is written as:

$$\bar{w}.x_+ + b \geq 1 \qquad (6)$$

$$\bar{w}.x_- + b \leq -1 \qquad (7)$$

For mathematical convenience, add another variable y_i such that it is +1 for positive sample and -1 for negative sample points. So, the (6) and (7) both becomes:

$$y_i(\bar{w}.\bar{x}_i + b) \geq 1 \qquad (8)$$

$$y_i(\bar{w}.\bar{x}_i + b) - 1 \geq 0 \qquad (9)$$

y_i would be zero for samples on margin planes & non zero for rest of the sample points. Width of the street is given by (10) as:

$$Width = (\bar{x}_+ - \bar{x}_-).\frac{\bar{w}}{||w||} \qquad (10)$$

For a positive sample (8) gives us result :

$$\bar{w}.\bar{x}_i = 1 - b = \bar{x}_+ \qquad (11)$$

And similarly for a negative sample (9) becomes:

$$\bar{w}.\bar{x}_i = 1 + b = \bar{x}_- \qquad (12)$$

From (11) and (12), (10) can be rewritten as:

$$Width = \big((1-b) - (1+b)\big).\frac{\bar{w}}{||w||} = \frac{2}{||w||} \qquad (13)$$

In SVM our main goal is to maximize the width of the street, which means minimizing ||w||. In this paper the data points are not linearly separable hence, we have used RBF (radial basis function) as kernel. Any function f which follows the property $f(x) = f(||x||)$ is called RBF.

V. EXPERIMENTAL RESULTS

Experiments have been performed on different samples of French and English language. For training of the system 10 speech samples of English and French each are taken. The training of the support vector machine is done by the feature vector i.e MFCCs extracted from the audio clip. The MFCCs extracted from English language clips are marked into class -1 and the MFCCs from French language clips are placed into class +1. Fig.3 shows the triangular filter bank used for feature extraction.

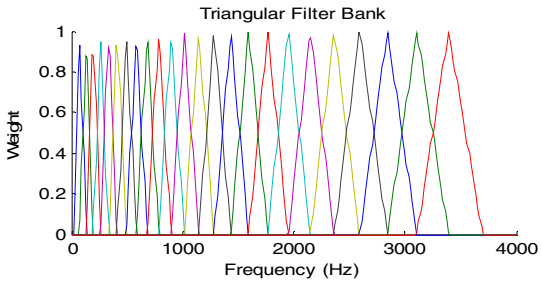

Fig 3: Triangular band pass filter

The spectrogram for one audio sample from English and French language is shown in Fig. 4.

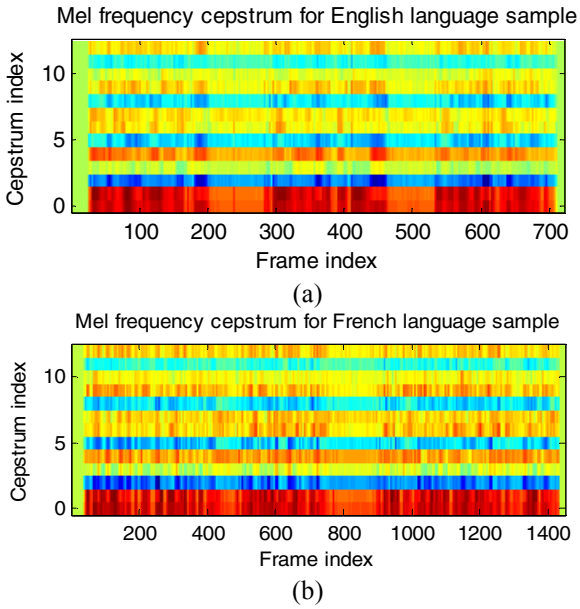

Fig 4 (a) Spectrogram for English language sample
(b) Spectrogram for French language sample

Now, the extracted MFCCs for English language are entered into a matrix data1 and for French the matrix is

data2. These matrices are used for SVM training. The plot of the data1 and data 2 is shown in Fig 5.

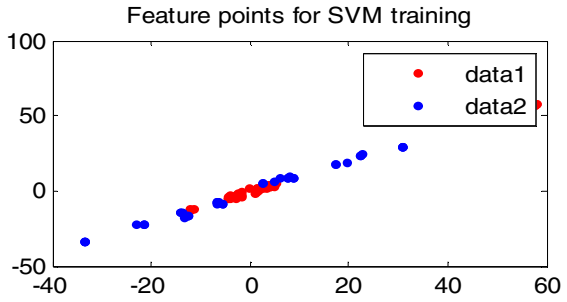

Fig. 5: Plot of MFCCs used for SVM training

The testing is done over 80 audio samples using RBF kernel. Fig. 6 shows the SVMs testing phase of the French audio clip.

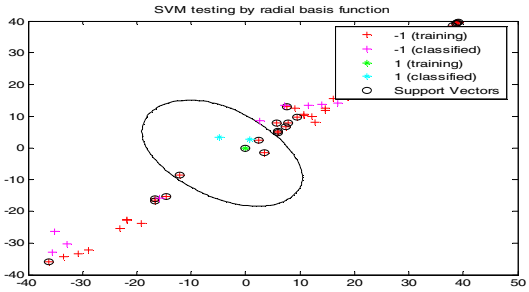

Fig. 6: SVM classification for French audio sample

Table 1 shows result of classification of some audio clips from the database created.

Table 1: Classification results for few audio clips

S.No	Audio Clip	Class Identified	Language Detected	Language Check
1	Eng_f1	-1	English	Correct
2	Eng_f2	-1	English	Correct
3	Eng_f3	1	French	Incorrect
4	Eng_m1	-1	English	Correct
5	Eng_m2	1	French	Incorrect
6	French_f1	1	French	Correct
7	French_f2	1	French	Correct
8	French_f3	1	French	Correct
9	French_m1	-1	English	Incorrect
10	French_m2	1	French	Correct

The accuracy for the proposed system is calculated as:

$$\text{Accuracy} = \left(\frac{\text{Total Number of Correct Samples}}{\text{Total Number of Testing Samples}}\right) \times 100$$

The same testing procedure is followed for many audio clips of English and French language. For English language, the proposed framework detects 35 clips correctly out of 40 testing clips giving an accuracy of 88 %. For French language, the trained framework identifies 37 samples correctly out of 40 testing clips giving a high accuracy upto 92%.

VI. CONCLUSION

In this paper, the automatic language identification between English and French language is achieved by using supervised support vector machine classifier with radial basis function. The SVMs are trained by the robust MFCCs extracted from the speech samples of English and French languages. The proposed scheme gives encouraging results with accuracy up to 88% and turned out to be better as compared to unsupervised classifiers. In future, the system can be expanded to discriminate between more than two languages by amalgamating two or more audio features.

REFERENCES

[1] D O'shaughnessy, "Acoustic analysis for automatic speech recognition" IEEE proceedings Vol. 101 Issue 5, 2013, page(s):1038-1053.

[2] W. M. Campbell, J. P. Campbell, G.A. Reynolds, E. singer, " Support vector machines for speaker and language recognition", Copmuter speech and language , volume 20, no. 2-3, 2006, page(s): 210-229.

[3] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk. "Speech Recognition using MFCC," International Conference on Computer Graphics, Simulation and Modeling , 2012, page(s): 135-138.

[4] William Campbell, Terry Gleason "Advanced Language Recognition using Cepstral and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation". In Proc. IEEE Odyssey, 2006, page(s): 1-8.

[5] W.M Campbel, "A SVM/HMM system for speaker recognition" in: Proceedings of the International Conference on Acoustics Speech and Signal Processing,2006, page(s):209–212.

[6] C.W Hsu , C. J Lin , "A Simple Decomposition Method for Support Vector Machines."Machine Learning, 46, 2002, page(s): 291–314.

[7] J. Weiner, N. T. Wu, D.Tellar, F. Maxeze, T. Schultz, D. –C, lyu, E. Chng and H. Li., " Integration of language identification into a recogination system for spoken conversations containing code switches" SLTU 2012, page(s); 76-79, 2012.

[8] M.W. Christopher, S. Khundanphur, and J.K. Baker, " An investigation of acoustic models for multilingual code – switching" In proceeding of INTERSPEECH, 2008, page(s): 2691-2694.

[9] Collobert, R., Bengio, S., SVMTorch: support vector machines for large-scale regression problems. Journal of Machine Learning Research 1,2001, 143–160.

[10] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, D. A. Reynolds. "Language Recognition with Support Vector Machines," MIT Lincoln Laboratory Lexington, MA USA.

[11] Smith, N. and Niranjan, M., "Data-dependent kernels in svm classification of speech patterns," in International Conference on Spoken Language Processing, 2000.