Indian Language Recognition

Vineeta Singh
Amrita School of Engineering

Deeptha Shree G.
Amrita School of Engineering

Abstract

Spoken Language Recognition is an important application with a growing necessity. The motivation behind this paper arises from the fact that India is home to 415 living languages (according to the SIL Ethnologue), *which highlights the necessity of an application which could identify each spoken language and classify it and further translate it.*
In this paper we investigate Indian spoken Language recognition for 3 languages namely Kannada, Tamil and Malayalam with the usage of RASTA-PLP for extracting features from speech utterances and Multi-layer Perceptron for classification of these features.

1. Introduction

Spoken language recognition is the process of identifying a speech utterance and classifying it. With the advent of globalisation, the demand for communication across boundaries is increasing. This has given rise to new challenges for Automatic Speech Recognition (ASR): before the machine can understand the meaning of the utterance, it must identify which language is being spoken. The past few decades has seen lot of advances in this field.
Humans have an inborn ability to distinguish and characterise languages to some extent. Scientists and researchers have been in a quest to automate this part of the human intelligence **[1].** Various ways of classifying a given language are acoustic phonetics, phonotactics, prosody and syntax.

- Acoustic phonetics: Phonemes are any of the minimal units of speech sound in a language that can distinguish one word from another. The number of phonemes in each language is between 15-50**.**Phonetic repertoire differs for different languages.
- Phonotactics: Each language has a particular set of phonotactics rules that determine the permissible phone sequences.
- Prosody: Prosody is the study of all the elements of language that contribute toward its acoustic and rhythmic effects.
- Syntax: The way in which linguistic elements (as words) are put together to form constituents (as phrases or clauses)

Based on these several methods have been devised for Language Identification (LID). However the basic steps involved in any LID remain the same:

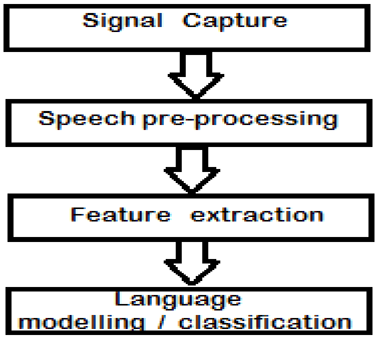

Fig1. Steps in language identification

2. STATE OF THE ART

This section discusses the background work related to different spoken Language recognition systems. This includes discussion on the algorithms which are most commonly used and the ones that have achieved the best results.

A lot of research has been performed in this domain with a lot of effort being put to improve the accuracy of the LID systems. Language identification is a two-step process: feature extraction and classification.

Various methods have been implemented for feature extraction. A few of them are described below:

- Linear predictive coding (LPC): It is a popular technique used for speech coding. It has the ability to provide accurate estimates of acoustic features with less computation and storage compared to other approaches. The LPC derives a compact and precise representation of the spectral magnitude for signals with a brief duration. The fundamental idea of LPC is that a speech sample can be estimated as a linear combination of past samples. One main limitation of LPC features is the linear assumption that fails to take into account of the non-linear effects. The LPC features also lack the discriminant power for classification tasks and are highly sensitive to acoustic environment. Additive background noise or room reverberation can affect the accuracy of LPC analysis.

- Mel-Frequency Cepstral Coefficients: The MFCC features apply Mel-frequency warping onto the power spectra with the use of a triangular Mel-scale filter bank. Logarithmic compression is also applied to Mel-spectra to approximate human auditory processing [5].

- Perceptual Linear prediction: PLP is an acoustically derived features proposed by Hermansky.[2] The following three concepts from the psychophysics of hearing are applied to derive an auditory spectrum estimate:
  - Critical-band spectral-resolution [6],
  - The equal-loudness hearing curve [7], and
  - The intensity-loudness power law of hearing [8].

- RASTA-PLP: A major cause of problem in speech recognition systems is the mismatch between the conditions used to record the speech training data and the conditions under which the data to be recognized is recorded (ex: a change in headset). The term RASTA comes from the words *RelAtive SpecTrA*. The RASTA technique applies a bandpass filter to each spectral component in the critical band spectrum estimate. Human hearing seems relatively insensitive to slowly varying stimuli [3]. The basic idea of RASTA filtering is to exploit these phenomena by suppressing constant and slowly varying elements in each spectral component of the short term auditory-like spectrum prior to computation of the linear prediction coefficients.

Feature vectors produced with these methods can be used directly in the training and testing of a vector-quantization-based, dynamic time- warping-based, hidden-Markov-model-based or neural network based language recognition system.

3. PROPOSED METHOD

The workflow of our proposed method is similar to figure1. The first step is to capture speech segments of Tamil, Malayalam and Kannada, the second step is pre-processing of the captured speech signal, the third step is feature extraction using RASTA-PLP algorithm and the fourth step is language identification with the help of MLP neural network.

3.1 SIGNAL CAPTURE

Speech segments of speakers in Tamil, Kannada and Malayalam were recorded using the open source software Praat. A total of 20 speech samples for each language were recorded from 3 different speakers in a quiet room. Each sample had a length of 3-5 sec.

3.2 SPEECH PRE-PROCESSING

Pre-processing of speech signals is considered a crucial step in the development of a robust language recognition system. It enables us to extract and represent useful speech information for the succeeding systems to work with the highest efficiency.

3.2.1 Silence Removal

Speech segments often contain areas of silence/noise which are useless for language recognition because they contain no information. Therefore, removal of noise would improve the efficiency of the system and reduce the computational complexity to quite an extent. To implement this we make use a certain threshold below which the signal is considered as noise and overwrite the original signal.

3.2.2 Normalisation

Normalisation is the process of equalizing the volume of audio files to a standard level. This is done because volume might vary from word to word in a speech segment.

3.2.3 Pre-emphasis

Pre-emphasis is done because speech samples distribute more in the lower frequency as compared to higher frequency. Pre-emphasis compensates for this suppression of energy in high frequency by the human vocal tract. It is done by passing through a first order high pass filter.

3.2.4 Framing/Windowing

We know that statistical properties of speech are not stable over a large period of time. Therefore, we opt for short time processing of speech signals. To perform this we used a hamming window because it has minimal side-lobes and smooth curve. This is implemented using the Enframe function in the Voicebox toolbox for Matlab. [4]

3.3 FEATURE EXTRACTION

In language recognition, feature extraction is given a lot of weightage because recognition performance depends heavily on this step. The main goal of the feature extraction step is to compute a set of feature vectors providing a compact representation of the given input signal. Through decades of research, many different feature representations of the speech signal have been suggested and tried. The most popular feature representation currently used is the Mel-frequency Cepstral Coefficients or MFCC.

Another popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. PLP was originally proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information

Feature extraction is performed on recorded speech segments of speakers in Tamil, Kannada and Malayalam.

To implement feature extraction we use RASTA-PLP. We use this technique to overcome the absence of robustness in the popularly used MFCC.
RASTA-PLP is used because it allows us to combat certain problems that we encounter in speech signal processing:

- Robust feature selection
- Mismatched additive background noise
- Mismatched input channels

RASTA-PLP is implemented using the signal processing toolbox for Matlab Voicebox [4].

PLP parameters obtained on applying RASTA-PLP to speech segments are compressed and given as input to the MLP neural network

4. LANGUAGE CLASSIFICATION

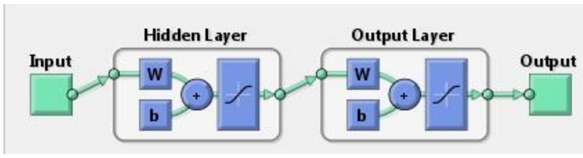

Fig2. Two layer Feed forward neural network

The compact PLP feature set of each speech utterance is given as input to a 2 layer feed forward Multi- Layer Perceptron model (MLP). Data is randomly divided for training, validation and testing. The artificial neural network is trained using scaled conjugate gradient back propagation algorithm. The performance is evaluated using mean square error and confusion matrices. Matlab Neural pattern recognition Toolbox was used to implement MLP feed forward network.

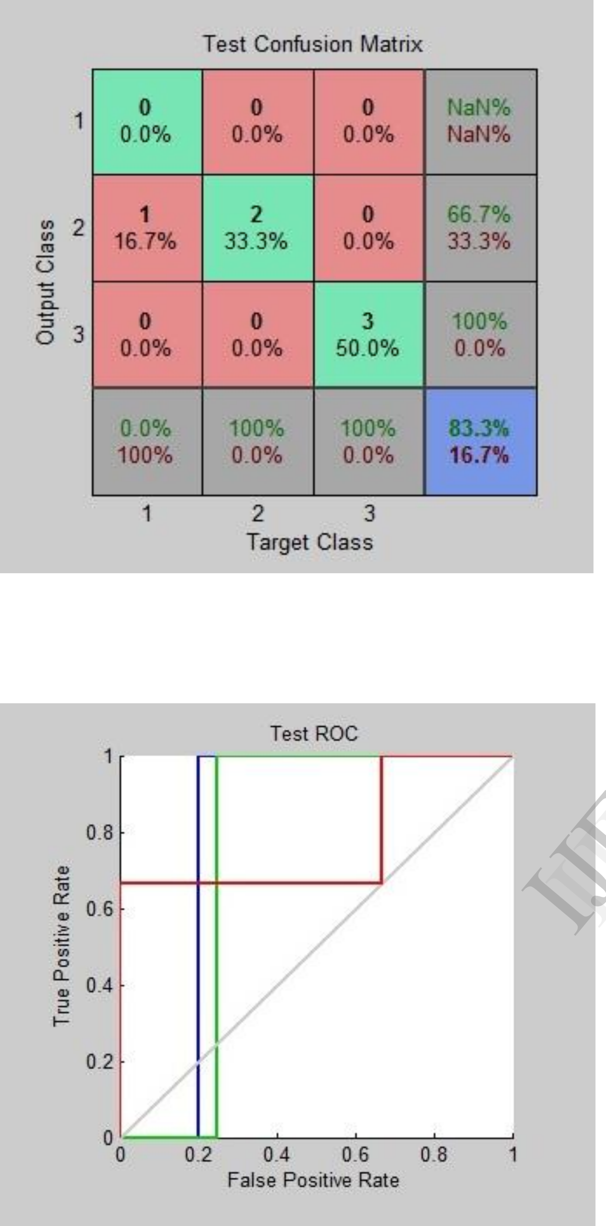

Fig3. Confusion matrix and region of convergence

5. CONCLUSION

The languages- Kannada, Tamil and Malayalam were classified successfully with good accuracy. Therefore, we conclude by stating that satisfactory results can be produced when Indian languages are classified with MLP feed forward neural networks using RASTA-PLP for feature extraction.

6. REFERENCES

1. J. Zhao, H. Shu, L. Zhang, X. Wang, Q. Gong, and P. Li, ‘‘*Cortical competition during language discrimination*,’’ NeuroImage, vol. 43, pp. 624–633, 2008

2. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

3. H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 578-589, Oct. 1994

4. {Ellis05-rastamat Author = {Daniel P. W. Ellis}, Year = {2005}, Title = {{PLP} and {RASTA} (and {MFCC}, and inversion) in {Matlab},Url = {http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/},Note = {online web resource}}

5. S. Davis and P. Mermelstein, “Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no.4, pp. 357–366, 1980.

6. Zwicker, E. “Masking and psychological excitation as consequences of ear’s frequency analysis,” in *Frequency Analysis and Periodicity Detection in Hearing* edited by R. Plotag and G. Smoorenburg (Sijthoff, Leyden, The Netherlands).

7. Makhoul, J. and Cosell, L., “LPCW: An LPC vocoder with linear predictive spectral mapping,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 466-469, Philadelphia, 1976

8. Stevens, S., “On the psychophysical law,” Psychol. rev. 64, 153-181