Language Recognition in Speech Audio

Harsh Manocha
2012134

Magus Verma
2012141

I. Problem Introduction

Many a times we come across audio/video content that is foreign to us by being in a different language. Such content becomes limited to the audience that can understand the language it is in. However, the content may appeal to anyone globally if only it was presented to them in language they could understand. Thus we believe a seemless translation system is what can solve the given problem. Though a lot of research work has taken place in speech recognition, but it has been mostly in speech-to-text conversion.

We believe in order to achieve the big aim of seemless translation of content across languages the first baby step required is to be able to predict the language of content with sufficient confidence. Here in this project we have contributed to solving of this problem by trying out different machine learning algorithms on various features and attributes that engineered from the audio data.

II. Dataset

There is no publicly avaialable dataset for this particular problem which has sentences from different languages spoken by multipe speakers. Since the problem is about predicting/classifying speech audio data, we dealt with audio files containing human speech in this project. Through a script, we downloaded dataset from TATOEBA. Our initial results were quite good on this dataset, but we later realized that a single language didn't have more than 2 speakers. So at the end of the day, instead of learning a language, we ended up learning the speakers.

To address this issue, we accumulated .wav extension audio files of 4 languages uploaded by about 50 distinct speakers in each language from websites like voxforge (http://www.voxforge.org/) using Javascripting, Python scripts and shell scripts. Finally, Each audio clip ranges in length from 5 to 7 seconds and had 48 KHz sampling rate. The bits per sample were 16.

Dataset consisted of following Languages that was used for formulating the classification problem.

- English - 500 Samples
- Spanish - 500 Samples
- French - 500 Samples
- German - 500 Samples

The data was ensured to have variety of speakers so that the machine learner we build does not merely learns the speaker but tries to understand the properties of a language in audio speech.

III. Preprocessing of Data

Generally the starting of the audio clip and ending has silence. The center part is more likely to contain words, and to be of better quality. We stripped the audio clip to preserve only 2 seconds at the middle (1 second before the center, and 1 second after). For preprocessing, we multiplied the signal with the hamming window to smoothen the signal.

The audio clip of 2 seconds having 48KHz sample rate is then divided into sliding windows of width 50 milliseconds, with the stride of 20 milliseconds, which produces 98 frames.

IV. Features and Processing

Audio data can be inherently decomposed into a mathematical combination of several sine waves of different frequencies with varying amplitude. This commonly is represented as the 3 dimensional spectrogram of the audio information and is used commonly in audio digital processing. This formation helps in deducing a lot of information about the entities that might be involved in a audio clip. In given problem different jaw movements during a speech generates different patterns.

Features constructed from the raw audio dataset for classification problem are as follows,

- MFCCs (Mel Frequency Cepstral Coefficients) : These helps in getting better representation of sound. Since it transforms the signals to a non-linear scale which corresponds to that of human ear, this representation turns out to be a useful feature.
- Spectral Centroid : Position of center of Gravity of spectrogram. The spectrogram captures the information on how each of different frequencies vary with time. Features related to spectrogram can help in telling what kind of sounds are there, and sometimes different languages have different sounds.
- Energy : Sum of squared signal values. This helps in capturing boldness of sounds, which might be different for each language.
- Entropy Energy : Distribution of Amplitude values in spectrogram.
- Spectral Spread : moment value of the audio sample spectrum.
- Zero Crossing Rate : Density of times a signal switches its value between positive and negative values. This captures

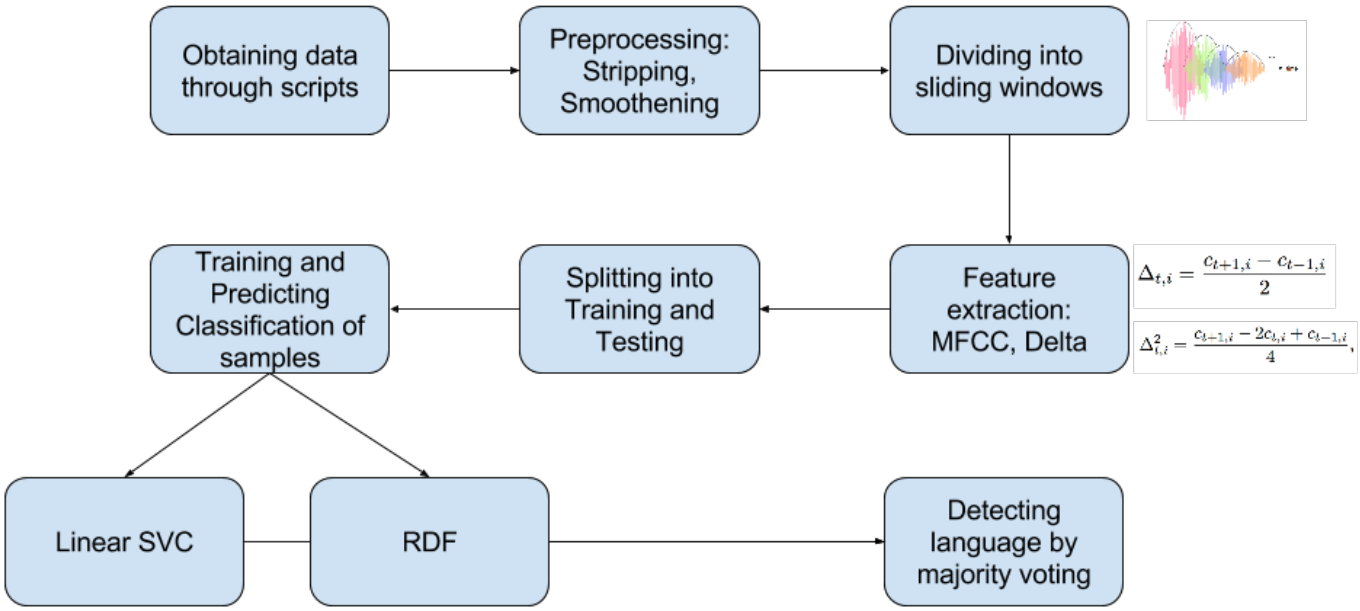

Fig. 1: Flow-chart showing summary of the approach

sounds like "fa" which involves different relaxations. Also, languages which have higher speed will tend to have higher ZCR.

- Since pace of the language is an important factor, we have tried to capture it via the first order and second order derivative of the mel-frequencies with respect to current frame.
  Delta 1 (First Derivative) : Taking the difference of cepstral values of t+1 for ith coefficient from t-1. 13 features for 13 cepstral values.
- Delta 2 (Second Derivative) : Taking the second derivative of cepstral values. 13 features for 13 cepstral values.

V. Models and Techniques

The audio can be represented as a simple linear combination of waves. This motivated us to attempt a linear classifier. On applying Linear SVM on the dataset with the mentioned features we found the accuracy to be not being much different from a random classifier.

We then looking at the uniqueness of spectrograms across languages intuitively felt that the boundary for the given problem is quite complex and looking at the result of linear classifiers we need a higher dimension kernel classifier or a different model that could produce a non linear boundary.
Audio data consists of values that are sampled as much as 48000 values for even a second of data, our assumption of sliding window thus returns as much as 98 datapoint for a second of data. Looking at high order of training values, we later realized on experimentation and theoretical digging that an rbf kernel svm sort of algorithm cannot produce a result in sufficient time.
Finally, we used Random Decision Forest with tree size of 100 and obtained quite better accuracy results in prediction. The classifier achieved accuracy achieved as much as 90

VI. Results

Classifier	Accuracy (Hamming)	Accuracy(no Hamming)
Support Vector Machine (Linear Kernel)	32.75%	43.25%
Random Decision Forest	69.50%	71.75%

TABLE I: Language 1 Vs 1 Classifier Results

	ENGLISH		FRENCH		GERMAN		SPANISH	
	SVM	RDF	SVM	RDF	SVM	RDF	SVM	RDF
ENGLISH			50%	85%	54.5%	89%	54.5%	81%
FRENCH	50%	85%			50.5%	84.5%	62%	78%
GERMAN	54.5%	89%	50.5%	84.5%			50%	84.5%
SPANISH	54.5%	81%	62%	78%	50%	84.5%		

VII. Conclusion

The problem is hard and requires a large data set and in-depth knowledge of signals and system, and the linguistic knowledge.
The preprocessing of multiplying with hamming window did not result in increase of accuracy. In fact, on the contrary, decreased it. This is probably because non-smoothned signal has information which can distinguish one language from another. RDF performs considerably better than linear SVM, which tells us that the data is not linearly separable. We tried applying RDF kernel based SVM, but that required extremely

large computation time since there are about 2 lacs samples. We also tried GMM (Gaussian Mixture Model), but that did not give good results (very near to a random classifier). Had the dataset been better (not so noisy, and speakers were not monotonically speaking), the results would have been better. None the less, this system tells us that it is indeed possible to recognize language automatically from an audio speech. This can be useful in seemless translation, and in applying automatic speech-to-text.

REFERENCES

[1] Vikramjit Mitra1, Daniel Garcia-Romero2, Carol Y. Espy-Wilson, *LANGUAGE DETECTION IN AUDIO CONTENT ANALYSIS*, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference

[2] https://github.com/tyiannak/pyAudioAnalysis, pyAudioAnalysis

[3] Shyamal Buch, Jon Gauthier, Arthur Tsang, X Li, *Language identification and accent variation detection in spoken language recordings*, Standford University