MULTILINGUAL SPEECH IDENTIFICATION USING ARTIFICIAL NEURAL NETWORK

Sweety Sarma[1]and Anupam Barman[2]

[1]IT.Developer, Assam Electronics Development Ltd
[2] Sr.System Consultant, Assam Electronics Development Ltd.

ABSTRACT

Speech technology is an emerging technology and automatic speech recognition has made advances in recent years. Many researches has been performed for many foreign and regional languages. But at present the multilingual speech processing technology has been attracting for research purpose. This paper tries to propose a methodology for developing a bilingual speech identification system for Assamese and English language based on artificial neural network.

KEYWORDS

Speech identification, cepstral analysis, artificial neural network, back propagation.

1. INTRODUCTION

Humans interact with others effortlessly using speech. The speech of a person differs and varies from one another with respect to some variables such as age, gender, emotional state, background noise, ascent, pronunciations etc.[1]

Speech recognition is an integral part in human computer interface. The speech technology is an emerging technology which aims to take a speech signal as input and detect the uttered word and produced it as an output.

Although various works has been performed for multi lingual ASR in foreign languages, much work has not been done with respect to Indian regional languages for multilingual purpose. C-DAC, as a partner of Asian Speech Translation Advanced Research (A-STAR) consortium aims to demonstrate speech translation from other regional languages to Hindi in the traveling and tourism area. Using Multi-lingual Speech data for 6 Indian different languages namely Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu and Indian English considered as seventh language approaches has been made in this aspect.[1]

This emerging technology is useful in many tasks such as applications on secure telephony, voice authentication system etc.[1][2]

2. LITERATURE REVIEW

Fuliang Weng, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke in their paper, “A Study on Multi lingual Speech Recognition”, describes their work in developing a multi lingual speech recognition system with special reference to English and Swedish language. They realised the acoustic component of the multilingual speech recognition systems by sharing Gaussian codebooks across the Swedish and English allophones. The language model was constructed by

DOI:10.5121/ijitcs.2015.5401

training statistical bigram model and a common backoff node. They had combined two monolingual language models into a probalistic finite state grammar and sharing parameters across two languages outputs better performance which is proved by their experimental results.[3]

Luk´aˇs Burget, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondˇrej Glembek, Nagendra Goel ,Martin Karafi´at, Daniel Povey, Ariya Rastrow, Richard C. Rose, Samuel Thomas, in their paper, "Multilingual Acoustic Modelling for Speech Recognition Based On Subspace Gaussian Mixture Models", presents experiments on a different approach on multi lingual speech recognitions, in which the phone sets are entirely distinct but the model parameters are not tied to specific states but are shared across the languages by using a model known as "Subspace Gaussian Mixture Model" where states' distributions are Gaussian Mixture Models with a common structure, constrained to lie in a subspace of the total parameter space over a very small amounts of in very small amounts of language training data. They intend to improve recognition performance for one language by training the shared parameters of the acoustic model on data from other language . Their experimental results shows that when the amount of training data for the target language is extremely limited, extremely large (Word Error Rate) WER reduction of 10.9% absolute by using data from other languages to train the shared parameters is achieved.

Thomas Nieslery, Daniel Willettz in their paper. "Language Identification and Multilingual Speech Recognition Using Discriminatively Trained Acoustic Models", depicted that they had performed language identification experiments on four different prominent South African languages using a multilingual speech recognition. Four South African languages are successfully recognised using single set of HMMs and a single pass recognition. The experimental results demonstrates that the use of discriminative training can improved the performance of language identification. Interestingly the authors did not constructed language models for their experimental purpose. [4]

Kim-Yung-Wong, in his thesis, "Automatic Spoken Language Identification Utilising Acoustic and Phonetic Speech Information", proposes a technique which produces more accurate and fast automatic spoken LID compared to previous National Institute of Standards and Technology (NIST) Language Recognition Evaluation. Wong employed Gaussian Mixture Model to model acoustic and phonetic features that are extracted using existing recognition components.[5]

Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze in their paper, "Multilingual Speech Recogntion", designed a single common framework together with language component that runs recognition engines in three languages viz; German, English and Japanese. Their experimental results shows a user friendly and flexible multi lingual spoken dialog system.[5]

Pinki Roy and Pradip K. Das in the research paper, "Language Identification of Indian Languages Based on Gaussian Mixture Models", designed a Language Identification Model that can successfully recognize four different Indian languages viz; English, Hindi, Assamese and Bengali. The classification was performed using Gaussian Mixture Model and evaluation was performed on a standard database . The experimental results shows that the system developed by them can give accuracy till the order of 1024.[6]

3. SPEECH IDENTIFICATION USING MFCC AND ANN

In general speech identification system, the pre-processed sound signal is fed to the feature extraction module and based on acoustic model the identification or recognition of the signal is done. There are various feature extraction techniques and classification techniques for speech processing. In this study, we have made an approach using MFCC as feature extraction techniques and artificial neural network as classification. MFCC technique is best technique for feature extraction.

A. MFCC

Among all the various feature extraction method MFCC is the best technique as per researchers suggest as it is considered to be the best available approximation of human ear. This technique generates the training vector by transforming the signal in frequency domain, so, it is less prone to noise. MFCC is achieved by several blocks like, sampling, pre-emphasis, windowing, fast fourier transform, absolute value, mel-scaled filterbank, log, discrete cosine transform, dynamic features, linear discriminant analysis etc. Thus it is best approximate to human ear.

B. ANN

An artificial neural network (ANN) is a kind of information processing method which simulates the working of biological nervous systems, such as the brain, process information and is basically based on the human nervous system. It comprises of the huge number of highly interconnected processing elements known as neurons or nodes to provide solution for specific problems. An ANN can be configured according to the specific application, such as data classification or pattern recognition, with the help of a learning process that can be specified as knowledge gaining process. An excellent feature of ANNs is that they can be trained easily using learning algorithm which can be used in area of research for various purpose. There are many types of neural networks which are generally used for pattern recognition process like Radial-basis function (RBF), Self Organising Map (SOM), Multilayer feed forward neural network etc. But most widely used neural network is Multi-layer feed forward Network using Back propagation algorithm. Multi-layer perceptron or multi-layer feed forward network consists of three layers where each layer give input to each next layer . The main idea is to propagate the information from layer to layer. A transfer function is attached which gives extra information. Back propagation algorithm is an iterative algorithm designed to minimise the mean square error between actual output and desired output. The algorithm iterates till the error is minimised or the difference between actual and desired output becomes minimum. [7][8][9][10][11][12]

4. IMPLEMENTATION ,RESULT AND DISCUSSION

The experiment was performed under open platform. Raw speech signals were collected for experimental purpose. Male and female voices are recorded using laptop microphone in a noise free environment at 44.1 Khz. 20 mixed sentences were recorded with 10 speakers uttering each word 10 times. Thus as a total we have 2000 recordings of mixed word samples Among these 1500 are used as training and remaining as testing dataset. The necessary pre-processing are performed using Praat tool which is an open source tool. After performing all the necessary pre-processing steps of the voice samples, the feature extraction is done.

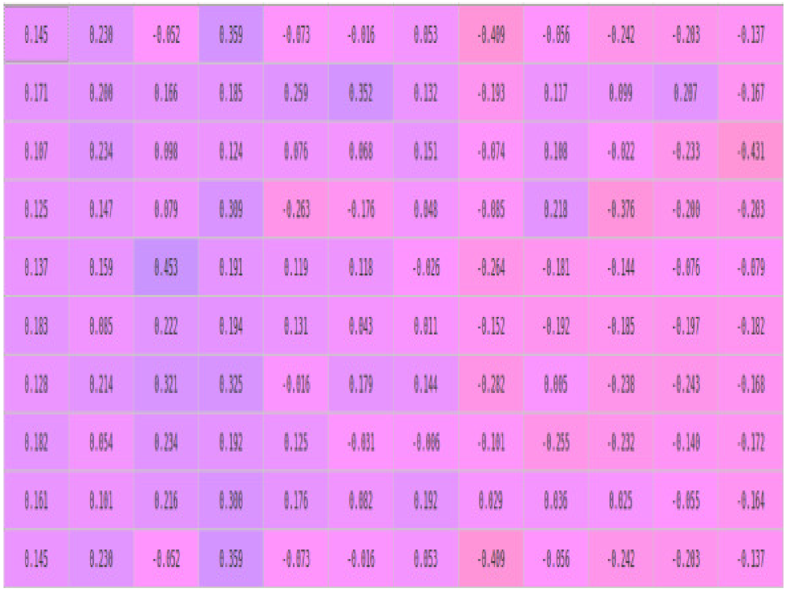
Some of our samples of language corpus are shown below:

Sl.No	Sentence
1.	মই এজন student
2.	মোক call কৰিবা
3.	কল এবিধ পুষ্টিকৰ fruit
4.	টমি এটা পোহণীয়া dog
5.	Bank কালি বন্ধ

6.	আজি shopping কৰিব যাম
7.	Library পৰা কিতাপ আনি লবা
8.	Finally কামটো হ'ল
9.	ৰাধা dance কৰে
10.	Dinner কৰি লোৱা

Table I: Language Corpus

The feature extraction is most important phase in developing a speech recognition system as it helps to differentiate between the utterances. It has been studied that good feature extraction results efficient output. In the experiment MFCC feature extraction technique is used as it is considered to be best approximation to human ear. Basically,10-12 coefficients are sufficient. 13 coefficients are calculated for our experimental purpose. The calculated set of MFCCs are used as inputs to the neural network.



0.145	0.230	-0.052	0.359	-0.073	-0.016	0.053	-0.409	-0.056	-0.242	-0.203	-0.137
0.171	0.200	0.166	0.185	0.259	0.352	0.132	-0.193	0.117	0.099	0.207	-0.167
0.107	0.234	0.098	0.124	0.076	0.068	0.151	-0.074	0.108	-0.022	-0.233	-0.431
0.125	0.147	0.079	0.309	-0.263	-0.176	0.048	-0.085	0.218	-0.376	-0.200	-0.283
0.137	0.159	0.453	0.191	0.119	0.118	-0.026	-0.264	-0.181	-0.144	-0.076	-0.079
0.183	0.085	0.222	0.194	0.131	0.043	0.011	-0.152	-0.192	-0.185	-0.197	-0.182
0.128	0.214	0.321	0.325	-0.016	0.179	0.144	-0.282	0.005	-0.238	-0.243	-0.168
0.182	0.054	0.234	0.192	0.125	-0.031	-0.006	-0.101	-0.255	-0.232	-0.140	-0.172
0.161	0.101	0.216	0.300	0.176	0.082	0.192	0.029	0.036	0.025	-0.055	-0.164
0.145	0.230	-0.052	0.359	-0.073	-0.016	0.053	-0.409	-0.056	-0.242	-0.203	-0.137

Fig I: Feature Extraction output

We have used a classifier to differentiate between the languages for words i.e set a threshold for classification.

The implementation of our work is done using Python. Calculation of MFCCs and creation of neural network training and testing is done using Python. ANN is used in our experiment as this classification method gives high accuracy rate. Moreover, in survey it has been found that speech identification is mostly done using Gaussian Mixture Model as the classifier.

The training of neural network is very important in artificial intelligence. Training is done to the neural network so that a particular input maps with the specific targets of output. 1500 samples are used as training set. And 500 samples along with few unseen sounds are used for testing purpose of the ann. **Multi layer perceptron neural network** is used in implementing neural network. **Back propagation algorithm** with momentum is used for training the designed neural network.

Parameter	Value
Learning Rate	0.2
Momentum	0.5
Sigmoid Function	1/(1+e^(-x))
Iteration	100000
Error	0.0

Table II: Parameters considered in designing ANN

We have successfully identified the words in both the languages (Assamese & English).Testing of the model is performed for both seen and unseen speaker i.e the one whose voice has been trained and whose voice has not been trained. Identification rate is high for known speaker while the rate is satisfactory for unknown speaker also.However, we have encountered false positives also for some ambiguous words.

For example, “call” and “কল” sounds same but both the words differ in language and meaning. “কল” word means “banana” in English which is a fruit. Since, the utterance of both the word is same so it becomes ambiguous and the model fails to recognise which word is actually pronounced. And it has been observed that sometimes it is identified as “call” and sometimes as “কল”. Moreover there are some words which are uttered both in English and Assamese language. Though we know they are English words we often used them in Assamese sentences. For example; “bank”, ”student”, ”dinner”, “lunch”, “address” etc.

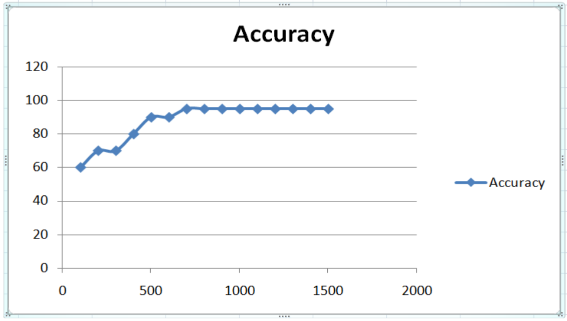


Fig II: Identification rate graph for seen voice.

Here, along x-axis we have plotted corpus. And along y-axis the identification rate.

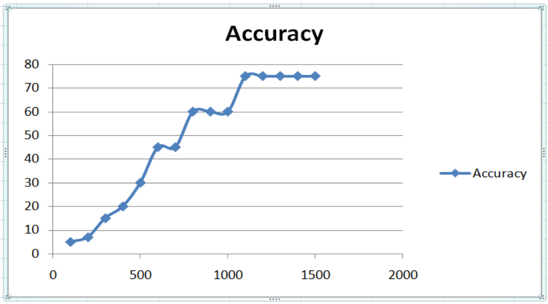


Fig III: Identification rate for unseen voice.

It has been observed that as the training set increases accuracy of detection also increases.
As the number of speaker increases the identification rate increases.

5. CONCLUSIONS

Speech processing is an important field in machine learning technology and multi lingual speech recognition is an emerging trend that has been attracting now a days. Various applications can be incorporate with this technology. Neural network is a significant tool in this process. Though some works are done for multilingual speech recognition, but most of them performed using Gaussian Mixture Model. Further we will try to work it in sentence level and using it in phoneme model.

In this paper, we have tried to implement a system in Assamese and English language using MFCC and artificial neural network and it has successfully identifies both Assamese and English words successfully according to their respective language. Moreover, neural network is useful classification tool in speech processing system.

REFERENCES

[1] G.Hemakumar & P .Punitha. “Speech Recognition Technology: A Survey on Indian Languages”, International Journal of Information Science and Intelligent System, vol.2, no.4,2013.

[2] Srivastava Nidhi, “Speech Reconition using Artificial Neural Network”, IJESIT, vol.3,2014.

[3] Fuliang Weng, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke , “A Study on Multi lingual Speech Recognition”, Speech Technology and Research Laboratory.

[4] Luk´aˇs Burget1, Petr Schwarz1, Mohit Agarwal2, Pinar Akyazi3, Kai Feng4, Arnab Ghoshal5, Ondˇrej Glembek1, Nagendra Goel6, Martin Karafi´at1, Daniel Povey7, Ariya Rastrow8, Richard C. Rose9, Samuel Thomas8, “Multilingual Acoustic Modelling for Speech Recognition Based On Subspace Gaussian Mixture Models”, 1 Brno University of Technology, Czech Republic, {burget,schwarzp}@fit.vutbr.cz; 2 IIIT Allahabad, India; 3 Boˇgazic¸i University, Turkey; 4 HKUST, Hong Kong; 5 Saarland University, Germany; 6 Virginia, USA; 7 Microsoft Research, Redmond, WA; 8 Johns Hopkins University, MD; 9 McGill University, Canada.

[5] Kim-Yung-Wong, “Automatic Spoken Language Identification Utilising Acoustic and Phonetic Speech Information”, Ph.D Thesis , Queensland University of Technology, 2004 .

[6] Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, Chin-Hui Lee, “A Study on Multilingual Acoustic Modelling for Large Vocabulary ASR”, ICASSP,2009.

[7] Pinki Roy and Pradip K. Das , “Language Identification of Indian Languages Based on Gaussian Mixture Models”, International Journal of Applied Pattern Recognition Journal ,vol.1,no.1,2013.

[8] Parsoya hiranjeev, “M.Tech Dissertation on Speech Recognition”,IIT Bombay,2014.

[9] Joe Tabelski, “Speech Recognition using Neural Networks”, School of Computer Science Carnegie Mellon University, 2001.

[10] Fiona Nielsen, “Neural Networks –Algorithms and Applications”, 2001.

[11] Urmila Shrawankar, “Techniques for feature extraction in speech recognition systems: A comparative study”.

[12] C.Santosh Kumar and Foo Say Wei, “A Bilingual Speech Recognition System for English and Tamil”, ICICS-PCM, 2003.