A Comprehensive Review of the Speech Dependent Features and Classification Models used in Identification of Languages

Chandrakanta Mohapatra
Pre-Phd Scholar
Dept. of Computer Application
North Orissa University,
Baripada, Orissa

Sujata Dash, PhD
Reader
Dept.of Computer Application
North Orissa University,
Baripada ,Orissa

Umakanta Majhi
Assistant Professor
Dept .of CSE,
NIT, Silchar
Assam

ABSTRACT

Automation of spoken languages become the need of the hour, and the advances in global communication have increased the importance of Language Identification, making feasible the availability of multilingual information services, such as checking into a hotel, arranging a meeting, or making travel arrangements, which are difficult actions for non native speakers. In this paper a comprehensive review of the approaches used in identifying spoken languages and the methods used for extracting speech dependent features are presented. In addition, different modeling techniques such as SVM, GMM, and PPRLM are reviewed, and how the change in speech feature characteristics can result change in the accuracy and performance of the system is also reviewed.

General Terms

Speech processing, Feature Extraction, Feature Classification, Classification modeling, Language Identification.

Keywords

LID-language Identification, SVM-Support vector Machine, GMM- Gaussian Mixture model, MFCC-Mel frequency cepstral co-efficient, PLP-Perceptual linear Prediction.

1. INTRODUCTION

Speech is one of the most natural and efficient means for communicating information among a group of people. Because speech communication is ubiquitous, researchers have made significant efforts to create methods for automatically extracting the fundamental information that a speech utterance conveys. Language identification (LID) is the process of determining the identity of the language corresponding to a given spoken utterance. The main task of automatic Language Identification (LID) is to quickly and accurately identify the language being spoken (e.g. English, Spanish, etc.) . Language identification has numerous applications in a wide range of multi-lingual services. Over the past four decades there have been continuous effort putted towards the automatic extraction of information from raw speech and using them to training and testing the classification models, for accurately recognizing the language spoken.

The structure of the paper is as mentioned ,Section 2 is a brief discussion about the previous work, and section 3 is about the review of the approaches and speech feature characteristics. Section 4 is the discussion about the result. Section 5 suggests implication and future research work.

2. SURVEY OF LINGUISTIC LITERATURES

Most language-ID systems operate in two phases, training and recognition or classification. During the training phase, the typical model or system is presented with examples of speech from a variety of languages, through that the system is trained. Some systems need only the digitized speech utterances and the corresponding true identities of the languages being spoken. More complicated language identification systems may require either (1) a phonetic transcription (sequence of symbols representing the sounds spoken), or (2) an orthographic transcription (the text of the words spoken) along with a phonemic transcription dictionary (mapping of words to prototypical pronunciation) for each training utterance. In order to build language models for each language, it is required to analyse the training speech through the generation of extracted features. The intent of these generated models is to represents the characteristics dependent on the languages taken into consideration for training. These models can then be used in the recognition phase of the language-ID process. During recognition, a new utterance is compared to each of the language-dependent models. [8][7]There is a variety of information that humans and machines can use to distinguish one language from another. At a low level, speech features such as acoustic, phonetic, phonotactic and prosodic information are widely used in LID tasks. At a higher level, the difference between languages can be exploited, based on morphology and sentence syntax.[9]

Acoustic information is generally considered as first level of analysis of speech production. Human speech is a longitudinal pressure wave and different speech events can be distinguished at an acoustic level according to amplitude and frequency components of the waves. Acoustic information is one of the simplest forms of information which can be obtained during the speech parameterization process directly from raw speech. Also higher level speech information such as phonotactic and word information can be extracted from the acoustic information.[7]

Once the basic acoustic features have been obtained, additional features are appended to each feature vector with the intention of incorporating the temporal aspects of the speech signal. Some commonly utilized additional features are the delta and acceleration cepstrum and the Shifted Delta Cepstrum. Phonotactics deal with valid sound patterns in a specific language, i.e. the allowable combinations of phonemes in a given language. The N-gram language model

(LM) can be used to model the phonotactic features. There is a finite set of meaningful sounds that can be produced physically by humans. Not all of these sounds appear in any given language and each language has its own finite subset of meaningful sounds .There is a wide variance in phonotactic constraints across languages. So, the phonotactic information carries more language discriminative information than the phonemes themselves and therefore it is suitable for exploiting the characteristics of a language. [6]

3. FEATURE EXTRACTION AND CLASSIFICATION MODELS

In the review, for each approaches it needs to extract the linguistic features from the speech inputs ,so need to generate the feature vectors prior to training the models and for testing it. The different features that used, are discussed below

The cepstral features contains the magnitude characteristic properties of the speech spectrum, and mostly used in language recognition process, where as the phonological features contains information about height of tongue, frontends of tongue, rounding of lip, nasalization, excitation, place and manner of articulation.[4]Again the prosodic features are also plays an important role for discriminating human speech. The information present in prosody are partially different from cepstral features as it contains information about tone, loudness, tempo or rhythm, so to utilize them effectively physical representation on features generation has to be carried out, these features include pitch, intensity and normalized duration of syllables. To extract the pitch information from every utterance of each language, the algorithm RAPT(Robust Algorithm for Pitch Tracking) can be used[4].

The mostly used feature parameterization technique in speech processing are PLP and MFCC ,in case PLP(perceptual linear Prediction) is based on psychophysics of hearing, which discards irrelevant information from the speech, so as it can improve the speech recognition rate, its spectral characteristics have been transformed to match the characteristics of human auditory system. This approximates three perceptual aspects namely the critical band resolution curve, the equal loudness curve & intensity loudness.[10]Where as MFCC is the computational algorithm which is realized by the bank of symmetric overlapping triangular filters spaced linearly in a Mel-frequency axis .This is accomplish by computing cepstral coefficients ,obtained by applying inverse DFT to the log energy output of the filter bank, in case of language identification the lower 12 co-efficients are used mostly for the cepstral feature vector.[6][4]

Different experiment carried out by the researchers shows, that there exists a slight difference in performance between PLP and MFCC .Some cases MFCC shows good performance over PLP, Other shows PLP having better performance than MFCC. So based on the number of co-efficient used ,the accuracy can be varied.[4] For the classification of feature vectors, in testing to hypothesize the language spoken ,the followings models are used

3.1 SVM (Support Vector Machine)

A SVM is a two class classifier ,it follows the one vs the rest strategy ,the main idea is to find a linear decision surface (hyper plane) that can separate one classes and has the largest distance (i.e., largest gap or margin) between border-line data points (i.e., support vectors). If such linear decision surface does not exist, the data is mapped into a much higher dimensional space (feature space) where the separating decision surface is found .The feature space is constructed via very clever mathematical projection (kernel trick),which is defined by the kernel function k(.,.) that maps the data in higher dimensional space.[3] The SVM framework is depends on choosing the appropriate kernel function which measures the distance or similarity between two sequences of speech feature vectors. W. M Campbell et al.[1] suggest this approach ,that uses the speaker and language identification using this classifier , which is a powerful technique for classification task ,here the main emphasis is given over the use of kernel that will compare the feature vector sequences and generates the similarity, for the recognition MFCC features are calculated and trained the model through SVM and used GLDS kernel, as in case of sequence kernel it is required to derive a function for comparing feature vectors, and through GLDS , SVM can be generalized to non-linear classifier to map features in higher dimensional feature space. The successfulness of SVM over other models is its use of kernel and second advantage is it is build upon a simpler mean square error classifier to produce more accurate result. Yan Deng, Jia Liu [3] used two approaches ,i.e support Vector Machine(SVM) and Phonetic N-gram, experimenting with two different ways of using SVM in the token based system, parallel phoneme Recognition, followed by language modeling, as in case of PPRLM there available m scores, for each input speech /utterance per target language , inorder to identify the language being spoken, requires a classifier and here SVM is used .Again Vicky and nitin Khanna[5]emphasized on the post processing of speech features before sending it to the classifier for classification, where they used k-mean clustering to reduce the huge number of MFCC features from speech .The proposed approach use k-means clustering which is one of the unsupervised algorithm, its used due to its simplicity and efficiency, this require to fix number of clusters k previously, and then starting with a random selection of k data points as K initial clusters, here the main aim is to minimize the sum of intra-cluster distance and maximizing inter-cluster distance, for the coverage of local minimum. here they build models through SVM ,for three language English,Hindi,Tibetian and achieve performance upto 81%.

3.2 GMM (Gaussian Mixture Model)

A GMM is a parametric representation of probalistic density function which is based on the weighted sum of multi variate Gaussian distribution. The training of a GMM involves through the formation of ,the estimation of probability distribution that best characterises the set of training data. Bo Yin1, and Eliathamby[2] gives an insight for the better performance of identification task by the speech inputs. Here the back end modeling was performed through GMM-UBM (Guassian Mixture model Universal background model)and by using the shifted delta cepstrum and feature warping technique. In the process of training the language models are adapted from the UBM using Bayesian adaption or MAP adaption. This is only applied to the mean of the mixture components, in contrast mean, weights and mixture.

3.3 PPRLM (Parallel Phone Recognition followed by Language Modelling)

In this case several single language phone recognition front ends are used ,in parallel to tokenize the input speech, then the output produced from phone sequences by the front ends are analyzed and a target language is hypothesized, the reason for the several single language phone recognizer is that ,the sound in the language to be identified is not always occur in one language.[3] Abhijeet Sangwan et al.[4] suggest the use of language feature from phonological features in feature extraction, additionally here also the pitch and energy based features are added, finally the proposed articulatory for language identification is combined with a PPRLM(parallel phone recognition language model),here the articulatory characteristics captured are height, front ,round, nasalization, excitation ,place ,manner of articulation, based on the movement of tongue.so each speech frame assumed to have seven articulatory values and here the language feature are used as if acoustic features used in modeling and classification, the modeling is performed based on the phonotactic analysis where multiple phone decoder used to tokenize each utterance before the classification.

4. RESULTS AND DISCUSSION

The evaluation was carried out by W. M Campbell et al.[1] to detect the presence of a hypothesized target language given a segment of speech. The target languages were American English, Egyptian Arabic, Farsi, Canadian French, Mandarin, German, Hindi, Japanese, Spanish, Korean, Tamil, and Vietnamese. Evaluation of the task was performed through standard measures: a decision cost function and EER. The training, development, and test data were primarily drawn from the CallFriend corpus available from the Linguistic Data Consortium (LDC). Training data consisted of 20 complete conversations (nominally 30 min) for each of the 12 target languages. Test data was consisting of length 3,10 ,30 sec., after building models through SVM it was tested ,and it is fusion with GMM approach to get the performance, the performance is listed in table no.1.

Bo Yin1, and Eliathamby[2] gives an insight for the better performance of identification task by the speech inputs, the features used for the system are not only the cepstral features but the combination of prosodic features. Here the back end modeling was performed through GMM-UBM(Gussian Mixture model Universal background model)and by using the shifted delta cepstrum and feature warping technique, there could improvement in accuracy by 87.1% on 10 language task, which outperforms the baseline system by nearly 12%,along with this ,through this paper author had researched to use the MFCC & PLP features of different co-efficients and compared the performance through a move in ,for better language identification task. Here the processing for the task is performed through training and testing .most commonly the different co-efficients used are 12 for MFCC and 9 for PLP,when training, the feature warping is used to normalize the distribution of features data to Gussian distribution and for the improvement of accuracy in the system. In contrast the comparision shows in the paper as MFCC-7 and PLP-7 co-efficients are the best co-efficient for average good performance in both PLP & MFCC i.e 77.8% and 78.2% respectively ,and the model’s overall accuracy is 87.1%.According to Yan Deng, Jia Liu [3] they had experimented on the conversational telephone database collected nearly 2000 conversations each having duration 1 to 20 minutes and five language in the group are used such as English ,Japanese, Korean, mandarin and Russian, here the testing is of 1500 segment,300 for each target and 30 sec speech signal, SVM Torch used for classification and backend processing is carried out through GMM. Again Abhijeet Sangwan et al.[4] suggest the use of language feature from phonological features in feature extraction, additionally here also the pitch and energy based features are added, finally the proposed articulator for language identification is combined with a PPRLM(parallel phone recognition language model),.Here they had used 5 south Indian language were taken for consideration i.e,Malayalam ,Kannada, Tamil, Telugu, Marathi, for modeling corpus consists of 75 hours of speech among this 65 hours was taken for training and 10 hours for testing purposes.

Later On Vicky and nitin Khanna[5]emphasized on the post processing of speech features before sending it to the classifier for classification,here k-mean clustering is used .Here MFCC is used for extracting feature from speech signal ,as it having frequency band equally spaced on non-linear mel scale and approximates human auditory system. They have extracted 24 MFCC features on every 30 mili second frame with frame shift of 50%,so the number of features for a speech signal of 1 min is 4000,as a result there is a sequence of feature vector for each speech input, in contrast a single feature vector so used the post processing algorithm before sending the features for classification, here after building model through SVM ,for three language English, Hindi, Tibetan, the table below shows the performances by the different approaches used by authors, and figure shows an easy visualization of the ERR in figure no.01.Prior to use of SVM, the use of GMM approach was performed better and when it is fusioned with backend fuser, it outperform both the approach. shows in the following table .

Table 1.Performance of different approach

Approaches Used	**EER(Expected Error Rate)**
SVM	6.1%
GMM	4.8%
Fused	3.2%
PPRLM	5.27%
PPRLM with SVM	3.72%

If we consider the accuracy performance through the use of approach and features used then the accuracy of identification for the hypothesized language is demonstrated in the table below.

Table.2. Accuracy of models based on features used

Approaches used	**Features used**	**Accuracy in %**
GMM-UBM	Cepstral feature & prosodic features	87.1%
PPRLM	Phonological features along with energy based & pitch features	86%
SVM	Post processed cepstral features	81%

5. CONCLUSION

As automatic language Identification task will be very helpful in recent days of communication, so it need to be accurately identify the spoken language ,it should not be baised towards a particular language and take less time to identify even the length of speech utterance can be less, so in this review it is observed that by using different algorithms and models like GMM and SVM and PPRLM we could get some accuracy and in some cases post processing of speech features could give better results ,but when this same modelling is tested over other language database that may not result in the same efficiency due to number of factors such as the human speech can be having varieties in age group, time of speaking, gender, environment factor, emotion, and the database will be very large, if consider all cases, and fitting into model over these data set captured will have different efficacy, so in a search of better accuracy over the identification, we could do research over the different ways followed and fusion of approaches can be researched in order to get some good accuracy.

6. ACKNOWLEDGMENTS

The authors would like to thanks the department of computer application, NOU ,Baripada, for the suggestion, support and assistance.

7. REFERENCES

[1] W.M. Campbell , J.P. Campbell, D.A. Massachusetts Institute of Technology, Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420, USA _ 2005 Elsevier Ltd.

[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] 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

[5] Vicky Kumar Verma and Nitin Khanna “Indian Language Identification Using K-Means Clustering and Support Vector Machine (SVM) Graphic Era University, Dehradun, Uttarakhand-248002,India(e-mail: Vickykverma7133@gmail.com). ©2013 IEEE

[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] Pedro A.torres-carrasquillo, Douglas A.Reynolds, J.R.Deller, ”Language Identification using Gaussian Mixture Model tokenization” Sponsored by DOD,US ©2002 IEEE.

[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.