Chapter 2

Literature review

This chapter provides an overview of existing language identification systems. In early 1970's, research in automatic spoken language identification was started. But, the progress in this area of research was slow for almost two decades. After that, with the advent of public-domain multi-lingual corpora for speech, many researchers started showing interest in this area and lot of effort and progress have been made [19][20]. All the existing LID systems use some amount of language-specific information either explicitly or implicitly and the amount of language-specific information used in them differs in both explicit and implicit LID systems. The performance and the complexity of the system are dependent on the amount of linguistic information supplied to the system and it is proportional. While training, some systems require only the speech signal and the true identity of the language. In these systems, language models are derived only from the speech data which is supplied during training. More complicated LID systems may require segmented and labeled speech signal of all the languages under consideration. Although the performance of the more complicated LID systems is superior to others. Adding a new language into more complicated LID systems is not a trivial task. Therefore, the trade off between performance and simplicity has become inevitable if the number of languages under consideration is large. A few representative LID systems in explicit and implicit groups are described in this chapter.

This chapter is organized as follows: An overview of explicit language identification systems is presented in section 2.1. Section 2.2 describes the existing implicit language identification systems. Importance of implicit LID systems over explicit LID systems are

given in section 2.3. Motivation for carrying out the present work is discussed in section 2.4. Summary and conclusions of this chapter is presented in section 2.5

2.1 Review of explicit LID systems

Muthusamy *et al* have done the perceptual experiments on language identification task and observed that the knowledge of a particular language (i.e, linguistic and syntactic rules) will definitely help in identifying or discriminating the languages [21]. From this, it can be interpreted that even for an automatic system, if speech recognizers of all the languages to be identified are used as front-ends, the performance in classifying the languages will be better. For developing a speech recognizer for any language, the basic requirement is, the segmented and labeled speech corpus.

Lamel and Gauvain used phone recognizers as front-end for language identification task [22][23]. Phone recognizers for the languages French and English are built and used in parallel. The unknown speech signal from any of these two languages is processed by the two phone recognizers in parallel. The language associated with the model having the highest likelihood is declared as the language of the unknown speech signal.

Berkling *et al.* [24], have considered a superset of phonemes of three different languages like English, Japanese, and German. They have explored the possibility of finding and using only those phones that best discriminate between language pairs.

Hazen and Zue [25] pursued a single multi-language front-end phone recognizer instead of language-dependent phone recognizer and incorporated the phonetic, acoustic, and prosodic information derived from speech within a probabilistic framework.

Andersen *et al.* [26] have grouped the total inventory of phonemes into a number of groups, one which contains language-independent phones and three language-dependent phone inventories for the three languages under consideration and tried to classify the languages.

Tucker *et al* [27] have utilized a single language phone recognizer to label multi-lingual training speech corpora, which have then been used to train language-dependent phone recognizers for language identification. They used three techniques such as the acoustic

difference between phonemes of each language, the relative frequencies of the phonemes of each language and the combination of above two sources of information for classifying the languages.

Zissman and Singer [28] used a single English phoneme recognizer and proved that it was feasible to model phonotactic constraints with the information of phoneme inventory from one language. Zissman [1] used a single language-dependent phone recognizer to convert the input speech signal to a sequence of phones and used the statistics of the resulting symbol sequences for language identification, which is termed as Phone Recognition followed by Language Modeling (PRLM). He has further extended it using multiple language-dependent phone recognizers in parallel (Parallel-PRLM) and achieved a reasonable improvement in the language identification performance.

Kadambe and Hieronymus [29] demonstrated that the performance of an LID system which is based only on acoustic models can be improved by incorporating higher level linguistic knowledge in the form of trigram and lexical matching. This system is also based on the Parallel Phone Recognition (PPR) approach.

Yan and Bernard [30] extended the Parallel Phone Recognition (PPR) approach which was discussed in [1] for six languages with refined bigram models and context-dependent duration models. For combining the evidence derived from the scores of acoustic model, language model, and duration model, they proposed a neural network based approach.

Navratil and Zhulke [31] have also used a single language-independent phone recognizer but have used improved language models instead of standard bigram models. The aim of using a single language-independent phone recognizer is to reduce the computation complexity introduced by the parallel phoneme tokenizers. Navratil improved the PPRLM LID system by using the binary-tree (BT) structures and acoustic pronunciation models instead of the traditional N-gram language models [32]. Two approaches of BT estimation are proposed - building the whole tree for each class in one case, and adapting from a universal background model (UBM) in the other case. The resulting system serves for language identification as well as for unknown language rejection, and achieved the error rates of 9.7% and 1.9% on the 1995 NIST (based on OGI-TS corpus) six-language identification task and 14.9% and 5.1% on the nine-language task for 10-sec and 45-sec test utterances respectively.

Hazen and Zue [33] proposed a phonotactic based LID system with a single decoder with a multilingual phoneme repertory and a variable number of phoneme units. The phonotactic classifiers use multiple phone recognizers as the front-end to derive phonotactic statistics of a language. Since the individual phone recognizers are trained on different languages, they capture different acoustic characteristics from the speech data. From a broader perspective, characterization of the spoken language is possible by combining these recognizers to form a parallel phone recognizer (PPR) front-end. GMM was applied to model the phonetic class in a segment-based approach. They achieved LID accuracy rates of about 50% measured on the 1994 NIST evaluation dataset, compared to about 70% achieved by a phonotactic component on the same data.

Kirchhoff and Parandekar [34] made an interesting study for modeling the cross-stream dependencies for the phonotactic based LID systems. In their approach, a multi-stream system was used to model the phonotactic constraints within as well as across multiple streams.

Prasad [35] developed a language-independent syllable recognizer for a two language task and tried using it as a front-end, in which the syllable statistics are used to determine language identity. Ramasubramanian *et al* [36] have studied the PPR system in detail for a 6-language task. They made a study on three different classifiers such as the maximum likelihood classifier, Gaussian classifier, and K-nearest neighbor classifier. For each classifier they have explored with three different scores, namely, acoustic score, language-model score, and joint acoustic-language model score and concluded that maximum likelihood classifier with acoustic likelihood score gives best LID accuracy.

Gauvain *et al.* [37] proposed another approach of generating a multitude of streams with the use of phoneme lattices. The use of phoneme lattices has been shown to significantly improve the performance of PPRLM systems when compared to the 1-best approach of considering only one phoneme (token) sequence [37, 38]. Gleason and Zissman [39] described two enhancements to the Parallel PRLM (PPRLM) system by using the composite background (CBG) modeling and score standardization.

In order to model reliably a longer time span than the traditional PPRLM (to model 5-gram instead of trigram), Cordoba *et al.* [40] presented an approach for language identi-

fication based on the text categorization technique. With the parallel phoneme recognizer as the front-end, the N-gram frequency ranking technique was used instead of the language model. The resulting LID system is capable of modeling the long span dependencies (4-gram or even 5-gram), which could not be modeled appropriately by the traditional N-gram language model, probably due to insufficient training data. The proposed parallel phoneme recognition followed by n-gram frequency ranking achieved a 6% relative improvement compared to the PPRLM LID system.

Li *et al.* [41] proposed to use a "universal phoneme recognizer", which was trained to recognize 258 phonemes from 6 languages (English, German, Hindi, Japanese, Mandarin and Spanish). For the back-end, both the N-gram models and the vector space modeling (VSM) were adopted to make a pair-wise decision. This PPR-VSM LID system achieved an EER of 2.75% and 4.02% on 30-sec test utterances for 1996 NIST LRE and 2003 NIST LRE respectively.

Sim and Li [42] improved the PPRLM based LID system by using the acoustic diversification as an alternative acoustic modeling technique. Unlike the standard PPRLM systems where the subsystems are derived using language dependent phoneme sets to provide phonetic diversification, the proposed method aims at improving the acoustic diversification among the parallel subsystems by using multiple acoustic models. By combining the phonetic and acoustic diversification (PAD), the resulting LID system achieved EERs of 4.71% and 8.61% on the 2003 and 2005 NIST LRE data sets respectively.

Tong *et al.* [43] proposed a target-oriented phone tokenizers (TOPT) that uses the same phone recognizer for different target languages in the PPR front end. For example, Arabic-oriented English phone tokenizer, Mandarin- oriented English phone tokenizer, as Arabic and Mandarin each is believed to have its unique phonotactic features to an English listener. Note that not all the phones and their phonotactics in the target language may provide equally discriminative information to the listener, it is therefore desirable that the phones in each of the TOPTs can be those identified from the entire phone inventory, and having the highest discriminative ability in telling the target language from other languages.

You *et al* [44] have used morphological information, including letter or letter-chunk N-grams, to enhance the performance of language identification in conjunction with web-based

page counts. Six languages, namely, English, German, French, Portuguese, Chinese, and Japanese are tested. Experiments show that when classifying four Latin languages, including English, German, French, and Portuguese, which are written in Latin alphabets, features from different information sources yield substantial performance improvements in the classification accuracy over a letter 4-gram-based baseline system. The accuracy increases from 75.0% to 86.3%, or a 45.2% relative error reduction.

Botha and Barnard [45] have exploited effects and the relations of different factors such as size of text fragment, amount of training data, classification features and algorithm employed, and similarity of the languages which affect the accuracy of text-based language identification. They have used 11 official languages of South Africa. Within these languages distinct language families can be found. They found that it is much more difficult to discriminate languages within languages families than languages in different families. They have used n-gram statistics as features for classification The relationship between the amount of training data and the accuracy achieved is found to depend on the windowsize: for the largest window (300 characters) about 400000 characters are sufficient to achieve close-to-optimal accuracy,whereas improvements in accuracy are found even beyond 1.6million characters of training data for smaller windows.

2.2 Review of implicit LID systems

The NIST language recognition evaluation conducted in the years 1996 [46] and 2003 [47], shows that the approaches based on a bank of parallel-phone recognizers of multiple languages were the best performing systems. But, some problems can be anticipated with these systems when the number of languages to be identified is increased. In [1], it is shown that, parallel-PRLM performs better than PRLM, in which a single phone recognizer is used as a front-end. Further, it is shown that reducing the number of channels (phone recognizers) has a strong negative effect in the performance. From this one can conclude that the performance of the Parallel-PRLM system is proportional to the number of speech recognizers used in parallel in the system. When the number of languages to be recognized is increased, the number of phone recognizers may need to be increased further. But developing a single phone recognizer

with reasonable performance itself is not a trivial task. This clearly shows that, attention should be given to developing language identification systems which do not require phone recognizers. The existing implicit LID systems differ mainly in the feature extraction stage, since the type of feature selected for discriminating languages may be different. Some of the representative implicit LID systems in the literature are discussed below.

The earliest sustained effort in automatic spoken LID systems were reported by Leonard and Doddington [48] at Texas Instruments (TI) labs. House and Neuburg [49] have grouped all the speech samples into five broad linguistic categories, with the assumption that it is possible to identify gross linguistic categories with great accuracy. They proposed the first HMM based language identification system. Cimarusti and Eves [50] showed that a pattern classifier approach can be used for language identification. They ran the experiment using a 100 dimensional LPC derived feature vector. Eady [51] performed a two language identification task (English and Mandarin Chinese) by examining the differences in the F_0 patterns.

Ives [52] extended the previous study by developing a rule-based LID system using an extended multi-lingual corpus. Foil [53] examined both formant and prosodic feature vectors and found that formant features were generally superior. The formant vector based language identification system used k-means clustering. Goodman and *et al* [54] have extended Foil's work by refining the formant feature vector.

Muthusamy *et al* [55] have suggested a segment based approach to identify the language, where the speech is first segmented into seven broad phonetic categories using a multi-language neural network based system with the basic idea that the acoustic structure of any language can be estimated by segmenting the speech into broad phonetic categories.

Sugiyama [56] performed vector quantization classification on acoustic features such as LPC coefficients, autocorrelation coefficients and delta cepstral coefficients. The difference between using one VQ code book per language versus one common code book was explored in [56]. In the latter case, languages were classified according to their VQ histograms. Riek [57], Nakagawa [4] and Zissman [58] used a Gaussian mixture classifier for language identification based on the observation that different languages have different sounds and sound frequencies.

Itahashi *et al.* [59], and Itahashi and Liang [60] proposed LID systems based on fundamental frequency and energy contours with the modeling using a piecewise-linear function. Li [61] proposed a LID system which is based on features extracted at syllable level. In this system, the syllable nuclei (vowels) for each speech utterance are located automatically. Next, feature vectors containing spectral information are computed for regions near the syllable nuclei. Each of these vectors consists of spectral sub-vectors computed on neighboring frames of speech data. Rather than collecting and modeling these vectors over all training speech, Li keeps separate collections of feature vectors for each training speaker. During testing, syllable nuclei of the test utterance are located and feature vector extraction is performed. Each speaker-dependent set of training feature vectors is compared to feature vectors of the test utterance, and most similar speaker-dependent set of training vectors is found.

Pellegrino and Andre-Abrecht [62] proposed an unsupervised approach to LID, in which each language vowel system is modeled by GMM trained with automatically detected vowels. Even though GMM for language identification is well experimented in [1] [63][64], the difference here is, the language models are generated using vowels alone. Corredor *et al.* [65], Dalsgaard *et al.* [66], Lamel and Gauvain [23], Pellegrino *et al.* [67], Ueda and Nakagawa [68], and Zissman [1] did extensive studies on LID using HMMs. Due to its abilities to capture temporal information in human speech, HMMs represent a natural bridge between the purely acoustic approaches and the phonotactic approaches [32, 1].

Cole *et al.* [3] applied ANN in the form of a multilayer perceptron trained by the PLP features. Braun and Levkowitz [69] described the use of the recurrent neural networks (RNNs) for the LID task. The GMM-UBM method was proposed for language identification by Wong *et al.* [5] and has gained momentum and become one of the dominant techniques for acoustic based language identification. Campbell *et al.* [70], Zhai *et al.* [6] and Castaldo *et al.* [71] applied SVMs for the language identification task and showed improved results compared to the GMM based approach. In a more recent development Noor and Aronowitz [72] combined the anchor models with the SVM.

In [63], Gaussian Mixture Model (GMM) is used to tokenize the speech signal and language models are derived from the sequence of tokens. Similar to phone recognizers in the

PRLM system, the GMM tokenizer is trained on one language but is used to decode information for any candidate language. The decoded sequence is then used to train bigram models. Further, a parallel GMM tokenization system has been tried, where for each language, a separate GMM tokenizer is used. It is shown that as the number of GMM tokenizers is increased beyond a level, there is a degradation in the performance. In this work Callfriend [20] corpus is used for both training and testing.

Lin and Wang [73] proposed the use of a dynamic model in ergodic topology with the input of the temporal information of prosodic features. Rouas *et al.* developed a LID system with only prosodic features, where a set of rhythmic parameters and fundamental frequency parameters were extracted. They later improved this with a modified algorithm of rhythm extraction and several prosodic parameters were extracted (consonantal and vowel durations, cluster complexity) and were modeled by the GMM [74]. The resulting LID system achieved a language identification rate of 67% on a 7-language task.

Chung-Hsien [75] *et al* have done segmenting and identification of mixed languages. A delta Bayesian information criterion (delta-BIC) is firstly applied to segment the input speech utterance into a sequence of language-dependent segments using acoustic features. A VQ-based bi-gram model is used to characterize the acoustic-phonetic dynamics of two consecutive codewords in a language. A Gaussian mixture model (GMM) is used to model codeword occurrence vectors orthogonally transformed using latent semantic analysis (LSA) for each language-dependent segment. They have achieved language identification accuracies of 92.1% and 74.9% for single-language and mixed-language speech, respectively.

Rouas [76] also implemented an LID system based on the modeling of the prosodic variations, which was achieved by the separation of phrase and accentual components of intonation. An independent coding of phrase and accentual components is proposed on differentiated scales of duration. Short-term and long-term language-dependent sequences of labels are modeled by n-gram models. The performance of the system is demonstrated by experiments on read speech and evaluated by experiments on spontaneous speech. An experiment is described on the discrimination of Arabic dialects, for which there is a lack of linguistic studies, notably on prosodic comparisons.

Siu *et al* [77] have used discriminative GMMs for language identification using boosting

methods. The effectiveness of boosting variation comes from the emphasis on working with the misclassified data to achieve discriminatively trained models. The discriminative GMMs aproach is applied on the 12-language NIST 2003 language identification task.

Sangwan *et al* [78] have done analysis and language classification based on speech production knowledge. The speech utterance is first parsed into consonant and vowel clusters. Subsequently, the production traits for each cluster is represented by the corresponding temporal evolution of speech articulatory states. Evaluation is carried out on South Indian Languages, namely, Kannada, Tamil, Telegu, Malayalam, and Marathi which are closely related. Accuracy rate of 65% was achieved with about 4 sec of train and test speech data per utterance.

Martnez *etal* [79] have extracted prosodic features such as intonation, rhythm and stress and classified the language with a generative classifier based on iVectors.

In the context of Indian languages, very few attempts are reported in the area of language identification. First attempt has been made on Indian languages by Jyotsna *et al.*, [80]. As Sugiyama, they also performed VQ classification on four Indian languages using 17 dimensional Mel Frequency Cepstral Coeffcients (MFCC). They observed that the acoustic realization of the same sound unit from different Indian languages are different and some sound units (key sounds) contribute to the performance of the LID system. Nagarajan et. al., [81] have explored different code book methods for building LID system. Later using automated segmentation of speech into syllable like units and parallel syllable like unit recognition are used to build implicit language identification system. In [2], Sai Jayaram and *et al* have used a parallel sub-word unit recognition (PSWR) approach for LID, where the sub-word unit models are trained without using segmented and labeled speech data. In this work, first, the training utterances are segmented into acoustic segments automatically and clustered using K-means algorithm. After clustering, HMMs for each class are generated. The rest of the work is similar to the PPR approach discussed in [1] and [36]. It is claimed that the language identification performance for this system is almost the same as that of the system which uses phone recognizers in parallel [36]. The resulting PSWR LID system achieved an LID accuracy of 70% on a six-language task (based on OGI-TS corpus) for 45-sec test utterances. Mary et. al., [82][83] have explored spectral features with autoassociative

neural network models for language identification with varying durations of test speech samples.In their later work [84], they focused on prosodic (intonation, rhythm and stress) and syllabic features for language recognition.

2.3 Reasons for attraction towards implicit LID systems

The most successful approach to LID uses the phone recognizer of one language or several languages as a front-end. Zissman [1] has compared the performance of four approaches for LID and argued that LID systems using Parallel Phone Recognition (PPR) approach will outperform other systems. But the basic requirement for this approach is the availability of phone recognizers of all the languages to be identified. To develop a phone recognizer for any language, a segmented and labeled speech corpus is necessary. Building segmented and labeled speech corpora for all the languages to be recognized, is both time consuming and expensive, requiring trained human annotators and substantial amount of supervision [85]. The other approaches tried in [1], do not require segmented and labeled speech corpora for all the languages to be recognized. Phone recognition followed by language modeling (PRLM) system explained in [1] requires only phone recognizer for any one language. Since all the sounds in the languages to be identified do not always occur in a single language used in the PRLM approach, it seems natural to look for a way to incorporate phones from more than one language into a PRLM system [1]. This approach, which uses more than one language speech recognizers, is referred to as Parallel-PRLM. Parallel-PRLM approach requires that labeled speech be available in more than one language, although the labeled training speech does not need to be available for all, or even any, of the languages to be identified. The performance of the Parallel-PRLM approach seems to be better than PRLM approach. But, as mentioned in the previous Section, the performance of the Parallel-PRLM system for language identification is proportional to the number of phone recognizers used in parallel.

Even though the performance of the PPR approach or the Parallel-PRLM approach is

impressive, non availability of segmented and labeled speech corpora of all the languages to be recognized, makes the implementation of PPR approach, harder. In all the above mentioned approaches (PPR, Parallel-PRLM), phone recognizers of several languages need to be used as a front-end for LID. The speech recognizer which will be used as a front-end for LID, should be capable of handling two mismatches; the channel mismatch and the language mismatch. Even though both of these mismatches can be handled to certain extent, the performance may not be optimal. The worse scenario is the non availability of speech recognizers of any language. The difficulties in implementing LID systems which rely on speech recognizers as the front-end, or the non availability of any of the speech recognizers, makes implicit LID systems more attractive, even though the performance is slightly inferior to that of explicit LID systems.

2.4 Motivation for the present work

The phenomena of globalization has brought people together from different parts of India. However, one of the barriers to global communication is due to existence of several languages and different people speaks in different languages. There are several hundreds of mother tongues exist in India. Almost 30 languages are spoken by more than a million native speakers. For an effective communication among the people across the states, appropriate speech interface is required. Here, the basic goal of the speech interface is to automatically transform the speech from source language to the desired target language without loss of any information. For example, a person from Tamilnadu can communicate with a person in Kashmir in their respective mother tongues. Here, speech interface has to identify the source and target languages based on initial speech samples, and then it has to transform the speech from source to target language. Speech transformation may be carried out with the following sequence of speech tasks : (i) speech to text transformation, (ii) source text to target text conversion and (iii) synthesizing the speech from target text. Therefore for speech to speech transformation language identification has to be performed first.

In Indian context, there is no systematic study on identification of ILs. The existing LID studies dealt with only 4 ILs. At present, there is no standard speech corpus covering

majority of ILs. But, for initiating LID task on ILs, standard speech corpus covering all ILs is essential. Therefore, in this work we have developed the IL speech corpus covering 27 ILs spoken by majority community. Due to complexity in accessing the resources for all 27 ILs, we have attempted implicit LID on ILs. For exploring the features to represent language-specific information, most of the existing works are based on spectral features extracted using conventional block processing and prosodic features extracted from phrases. Since, all the ILs are originated from Sanskrit, they have lot of similar characteristics in various aspects and hence the difficulty in discriminating them using conventional features. Therefore, in this work we have explored spectral features extracted from PSA and GCR for representing the language-specific information. Similarly in view of prosodic features we have explored intonation, rhythm and stress (IRS) features at syllable and word levels in addition to global level prosodic features. For minimizing the confuseability among the languages, complementary evidences from various features are combined at various phases.

2.5 Summary and conclusions

This chapter summarizes the existing philosophies in LID. Existing works on explicit and implicit LID are discussed in terms of features and models for capturing the language-specific information and fusion techniques for improving the performance. Tendency of present research on LID towards implicit LID task is briefly discussed. Existing works on LID in Indian context are briefly discussed. Finally the motivation for the present work is described.