UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Technical Report SU-EE-1201

A literature review of language, dialect and accent identification

Herman Kamper and Thomas Niesler

Digital Signal Processing Laboratory
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
kamperh@sun.ac.za, trn@sun.ac.za

19 January 2012

Abstract

This document gives a brief overview of literature dealing with language, dialect and accent identification. In the last two decades various researchers have attempted these identification tasks for numerous languages, dialects and accents and several techniques have been developed. This document aims to provide a brief overview of the most notable of these contributions. Approaches to language identification can roughly be divided into two groups: acoustic modelling, where spectral features of different languages are modelled directly; and phonotactic modelling, where speech is tokenised into phone strings and scored using different language models. Techniques from both philosophies are discussed. Many of the techniques developed for language identification are directly relevant to the dialect and accent identification problems and research dealing explicitly with these tasks are also discussed. From the given literature review we conclude that, for a specific language, dialect or accent, the most appropriate modelling technique is dictated mostly by the availability of data. If no transcribed speech data are available, acoustic modelling approaches seem most appropriate. If transcribed data for languages different from the target languages are available, approaches based on phone recognition followed by language modelling can be implemented. Where transcribed speech data for the target language are available, parallel phone or word recognition could be considered.

1 General Background

1.1 Evaluation of Identification Systems

Various measures can be employed for evaluating and comparing language identification systems.[1] Therefore, before a summarised review of the literature dealing with such identification systems can be given, a brief overview of these evaluation measures is necessary.

The language identification problem involves determining the language of a speaker from a given speech utterance. Perhaps the simplest measurement used to evaluate such systems is the identification accuracy, which is simply the number of correct identifications made by a system divided by the total number of utterances in the test set. The identification accuracy, or simply accuracy, is normally expressed as a percentage. A related measure, the recognition error rate (RER), gives the percentage of misidentifications for a given test set. Both these measures can also be given as part of a confusion matrix [1].

Many identification problems are considered a two-class prediction (binary classification) problem. For example, in speaker recognition we would often like to determine whether a given speech utterance contains speech from a specific speaker or not. Similarly language identification can be seen as a problem where we would like to determine whether a specific language is used in a utterance or not, i.e. a detection system. This normally involves using a statistical model to determine the likelihood of a given utterance containing some language and then confirming or rejecting that the language is present based on some likelihood threshold. When considering the language identification problem in this way there are four possible outcomes as illustrated in Table 1.

The probabilities of the two types of errors shown in Table 1 (false positives and false negatives) are often used for the evaluation of language identification systems. The two probabilities can be written as:

$$P(\text{false negative}) = P(\text{miss}|\text{target}) \tag{1}$$

$$P(\text{false positive}) = P(\text{false alarm}|\text{non-target}) \tag{2}$$

By performing identification of an evaluation test set, these statistics can be calculated and used to evaluate an identification system. Different operating points corresponding to different likelihood thresholds for separating true or false decisions can be considered. By varying the detection threshold the false negative (miss) rates can be plotted against the false positive (false alarm) rates resulting in so-called receiver operating characteristic (ROC) curves. For these curves the two error rates are plotted on linear axes and the optimal point (with zero identification errors) is at the origin. A variant of the ROC curve often used is the detection error trade-off (DET) curve. A DET curve is plotted with its axes on a normal deviate scale and therefore produce approximately linear curves [2]. One point of interest on ROC and DET curves is the equal error rate (EER) at which

$$P(\text{miss}|\text{target}) = P(\text{false alarm}|\text{non-target}) \tag{3}$$

EER is also often used to compare and evaluate different language identification systems.

By using ROC curves, DET curves and EER as evaluation measures, language identification is inherently considered a detection task comparable to that of speaker recognition. Some researchers may argue that this is inappropriate for the language identification problem, as

[1] In this general background section we are using 'language identification' to refer to either language, dialect or accent identification.

Table 1: Possible outcomes when viewing language identification as a detection task.

		Language actually present	
		true	false
Recogniser output	true	true positive	false positive
	false	false negative	true negative

discussed in [3]. Nevertheless, these measures are often employed for the evaluation of language identification systems, for example in the NIST language recognition evaluations described in Section 1.3.

1.2 Common Databases

Several standard databases have been used by different authors conducting research dealing with language identification. Some of these common databases are described in the following.

1.2.1 The OGI Multi-Language Telephone Speech Corpus

The Oregon Graduate Institute multi-language telephone speech (OGI-TS) corpus has been used extensively for comparison and evaluation of language identification systems. The corpus consists of speech collected in 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The training partition of the corpus consists of a total of about 90 minutes of speech from about 50 different speakers for each language. The development partition consists of messages from approximately 20 speakers in each language. Each speaker contributed between one and two minutes of speech, with no speaker speaking more than one message or more than one language. Phonetic transcriptions for six of the languages (English, German, Hindi, Japanese, Mandarin and Spanish) are also available [4, 5, 6, 7].

1.2.2 The Linguistic Data Consortium's CallFriend Corpus

The CallFriend corpus from the Linguistic Data Consortium consists of untranscribed telephone conversations in 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese (the 11 OGI-TS languages and Arabic). Data are provided for two dialects of English, Mandarin and Spanish, while only a single database for the remaining languages are provided resulting in 15 databases in total. The corpus is divided into training, development and evaluation sets. From [8] it appears as if all three sets consist of 20 telephone conversations lasting between 5 to 30 minutes each, in all 12 languages and the additional three dialects. For the training sets the conversations seem to be complete 30 minute conversations. No speaker occurs in more than one conversation [6, 9].

1.3 NIST Language Recognition Evaluations

In 1996, 2003, 2005, 2007 and 2009 the National Institute of Science and Technology (NIST) facilitated evaluations of automatic language identification systems. The aim of the NIST language recognition evaluations (LREs) is to determine baseline performance and state-of-the-art algorithms and techniques used for automatic language identification. Several sites take part in these evaluations and following each evaluation numerous papers describing the current state of

language identification are published. For all the evaluations speech utterances of three durations are considered: 3, 10 and 30 seconds. The task and the data made available change slightly from evaluation to evaluation and some of the details are given in the following.

1.3.1 1996 NIST LRE

The 15 languages and dialects included in the CallFriend corpus were considered target languages (the dialects were treated as separate languages). Training data (used mostly for training phone recognisers or acoustic models such as GMMs) could come from any source, but the CallFriend corpus was specifically recommended. The phonetically transcribed partition of the OGI-TS corpus was also used by some sites for training phone recognisers in each of the six languages: English, German, Hindi, Japanese, Mandarin and Spanish. Development data (often used for training back-end classifiers) consisted of approximately 1200 segments from the CallFriend corpus for each language and each evaluation duration (3, 10 and 30 seconds) while the evaluation test set consisted of approximately 1520 segments for each language and duration [10, 11].

1.3.2 2003 NIST LRE

The 12 languages included in the CallFriend corpus were considered target languages. Similar to the 1996 LRE, training data could come from any source. The development data for the 2003 LRE consisted of the combination of the development and evaluation sets for the 1996 LRE. The new evaluation set consisted of 1280 segments for each of the three test durations in each of the 12 target languages [12].

1.3.3 2005 NIST LRE

The 9 target languages and dialects were: American English, Indian English, Hindi, Japanese, Korean, Mainland Mandarin, Taiwan Mandarin, Mexican Spanish and Tamil. Training data could again come from any source. Development data consisted of the combination of the development and evaluation data from the 1996 and 2003 LREs. The 2005 LRE evaluation data were collected by the Oregon Health Sciences University (OHSU) and consisted of 513 English, 143 Hindi, 365 Japanese, 314 Korean, 644 Mandarin, 259 Spanish and 193 Tamil segments of each evaluation duration [3, 13].

1.3.4 2007 and 2009 NIST LRE

The LREs following 2005 continued in a similar trend, although the number of target languages was significantly increased (26 languages and dialects for the 2007 LRE and 23 languages for the 2009 LRE). Again training data could come from any source and development data consisted of the development and evaluation data provided for previous LREs. One notable difference came with the 2009 LRE, where in addition to the conversational telephone speech used for previous evaluations, radio broadcast data from Voice of America (VOA) were provided for development purposes and broadcast speech data were included in the evaluation test sets [14, 15].

1.3.5 Evaluation of Identification Systems in the NIST LREs

As noted in Section 1.1, language identification can be considered a detection task and this is the view adopted for the NIST LREs. Evaluation consists of testing each test segment against each

target language. So for the 2003 LRE, there were 12 trials for each test segment, corresponding to the 12 target languages [9]. For each trial the system either confirms or rejects the hypothesis that the language under consideration is present. From these trials estimates of $P(\text{miss}|\text{target})$ and $P(\text{false alarm}|\text{non-target})$ in (1) and (2) can be obtained, ROC and DET curves can be produced and the EER for a system can be determined. In some cases, as in the 2005 LRE, the contribution to miss and false alarm probabilities were modified to ensure that the contribution is equal for each language instead of implicitly weighing the contributions by the number of test segments from each target language [3]. This may become an issue when the number of test segments differs significantly between the different target languages.

2 Spectral Feature Modelling

Approaches to language identification can be roughly divided into two groups: acoustic modelling where spectral features of different languages are modelled directly, and phonotactic modelling [6, 16]. The latter is described in Section 3 while the former is discussed in the following. A summary of all the literature discussed in this chapter is given in Table 4.

When modelling spectral speech features directly, Gaussian mixture models (GMMs) are often employed. In an early paper, Zissman compared GMMs with ergodic (fully connected) hidden Markov models (HMMs) for the language identification problem [17]. While a GMM is a static classifier (i.e. it does not take temporal characteristics into account), ergodic HMMs can model the sequential characteristics of a speech signal without the requirement of transcribed data.

Zissman employed tied GMMs as HMM observation PDFs, which means that all state observation PDFs share the same set of Gaussians with state-specific mixture weights. This was done in order to reduce the number of parameters that needed to be trained. For each language considered, both GMMs and HMMs were trained. For a given test utterance, the language corresponding to the model yielding the highest likelihood was taken as the identity of the language being spoken. Various languages and corpora were considered in different identification experiments (see Table 4). Surprisingly the GMMs and the ergodic HMMs (with up to 20 states) showed similar classification performance. Some reasons for this are discussed in [17], including that the amount of training data was insufficient for training ergodic HMMs able to model the additional temporal characteristics effectively.

3 Phonotactic Modelling

Following on from the work in [17], Zissman compared Gaussian mixture modelling to phonotactic modelling [5]. Phonotactic modelling normally involves the tokenisation of input speech into phone strings and the subsequent scoring of these strings using different language models (LMs), with an LM trained for each target language. Based on the LM scores a classification decision is made.

Three different phonotactic modelling approaches were considered by Zissman in [5]. The first is referred to as phone recognition followed by language modelling (PRLM) and is illustrated in Figure 1. For this type of system training utterances for each target language are recognised using a single phone recogniser (not necessarily trained on one of the target languages) to produce phone sequences which are then used to train n-gram LMs for each language. The LMs are trained directly on the recogniser's output, not on human-transcribed data. During recognition the speech utterance is tokenised using the phone recogniser and the language associated with the LM yielding the highest score is taken as the hypothesised language. The advantage of

this approach is that transcribed speech data for only one language need be available. In [5] a phone loop grammar was used in the front-end phone recogniser and bigram LMs were used for the phonotactic analysis. A back-end classifier can also be trained on the different LM scores although this was not done in [5].

The second phonotactic modelling approach is referred to as parallel phone recognition followed by language modelling (PPRLM) and is illustrated in Figure 2. This approach extends the PRLM approach by using multiple front-end phone recognisers instead of a single recogniser. Input speech is presented to a bank of phone recognisers which are trained on the target languages, a subset of the target languages, or a different set of languages altogether. For each recogniser, the phone strings are then scored using a bank of LMs. A PPRLM system can therefore be seen as separate PRLM systems (using different front-end phone recognisers) running in parallel. The different LM scores can then be presented to a classifier. In [5] the outputs of the separate PRLM systems were averaged in the log domain (multiplied in the linear domain) as if each PRLM system was operating independently and the LM yielding the highest average score was then used to hypothesise the language of the input utterance.

For the PRLM and PPRLM approaches phone recognition is followed by phonotactic analysis using n-gram LMs. These approaches are reasonable to follow when limited transcribed speech data are available or transcribed data are only available for languages different from the target languages. When transcribed data in all the target languages are available it becomes feasible to simply train separate phone recognisers, each integrated with its own LM, and run these recognisers in parallel during identification. In this way the phone sequence obtained by each recogniser is optimal with regard to the combination of both acoustics and phonotactics. This approach is referred to as parallel phone recognition (PPR). As illustrated in Figure 3 the likelihood scores for each recogniser can be presented to a classifier after recognition. In [5] the recogniser yielding the highest likelihood score was simply used to identify the language of a test utterance. Clearly the disadvantage of the PPR approach lies in its requirement of transcribed data for all the target languages.

Zissman compared the three phonotactic modelling techniques and GMM-based classification using the English, Japanese and Spanish partitions of the OGI-TS corpus, including the available phonetic transcriptions (see Section 1.2.1). The training partition of the corpus was used for training while 10 and 40 second speech segments from the development partition were used for evaluation. PPRLM and PPR showed similar performance, which is not surprising since phone recognisers for all three languages were used for both approaches. The GMM-based classification system performed worst. Additional experiments were performed to compare GMMs, PRLM and PPRLM when identifying 10 languages (the OGI-TS languages excluding Hindi). In these experiments PPRLM (using six phone recognisers trained on the OGI-TS phonetic

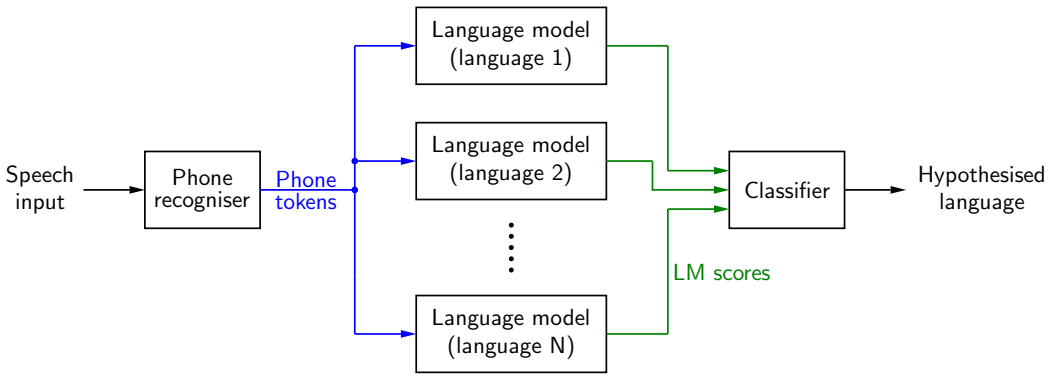

Figure 1: Language identification using phone recognition followed by language modelling (PRLM).

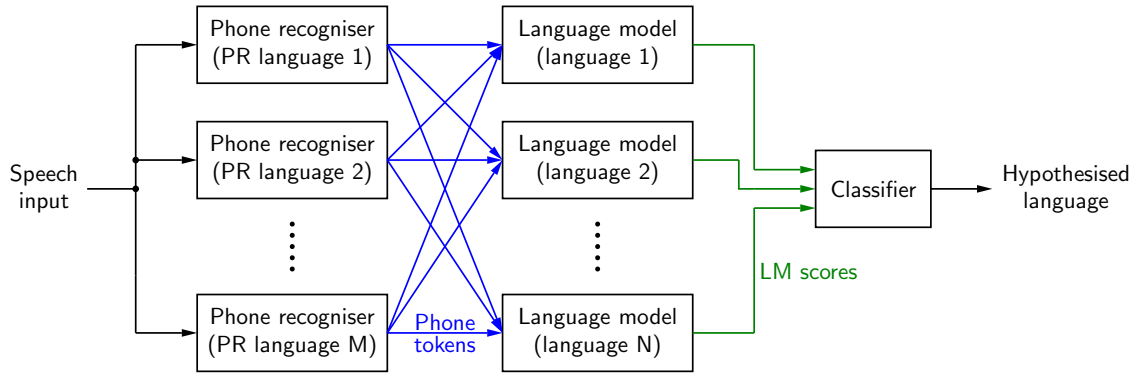

Figure 2: Language identification using parallel phone recognition followed by language modelling (PPRLM). The languages on which the phone recognisers are trained (PR language 1 to M) could correspond to the target languages 1 to N, could be a subset of the target languages or could be a different set altogether.

transcriptions) outperformed PRLM (using a English front-end phone recogniser) with GMMs again performing worst [5].

Several other authors followed language identification approaches based on variations of the PRLM, PPRLM and PPR techniques described above. For example, in [18] a PRLM system with an English front-end phone recogniser was developed for identification of the six phonetically transcribed OGI-TS languages. The authors obtained 57% and 78% identification accuracies on 10 and 45 second test utterances respectively. For the same languages, Yan and Barnard [19] obtained 81% and 92% classification accuracies for 10 and 45 second test utterances respectively. These authors implemented the PPRLM approach with a neural network back-end classifier. Fernández et al. [20] developed a PPRLM language identification system for English and Spanish as a baseline for comparison with other identification techniques. An identification accuracy of 92.6% was obtained using the PPRLM approach with the system employing both English and Spanish front-end phone recognisers.

4 Lexical Modelling

When appropriate data are available, language identification can also be accomplished as a by-product of large vocabulary continuous speech recognition (LVCSR). A few variants of the word-

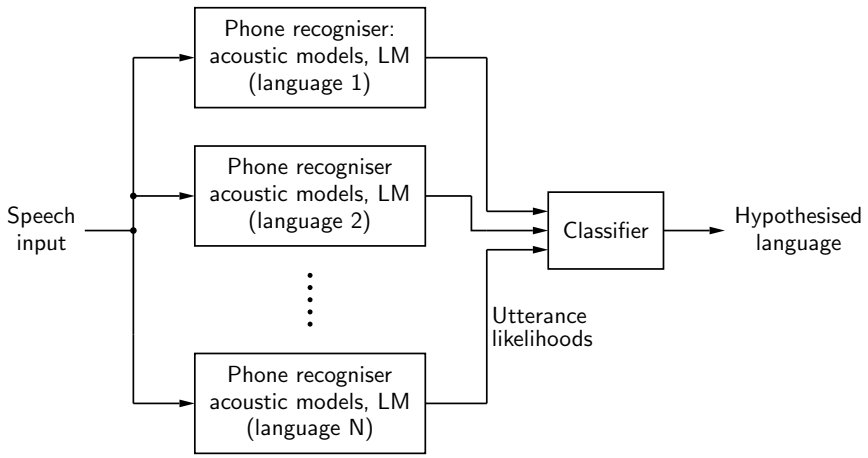

Figure 3: Language identification using parallel phone recognition (PPR).

based identification approach have been proposed. In analogy to PPR, Niesler and Willett [21] describe an approach were several word recognisers are run in parallel, each with its own acoustic model set, LM and pronunciation dictionary, as illustrated in Figure 4. These systems could be referred to as parallel word recognition (PWR) language identification systems. An alternative to this would be an approach that might be appropriately termed parallel word recognition followed by language modelling (PWRLM), similar to PPRLM but employing front-end word recognisers and performing back-end language modelling.

Schultz et al. [22] compared PPR, PWR and PWRLM in identification experiments for English, German, Japanese and Spanish. Two PPR systems were developed: the first employed phone recognisers with phone loop grammars while the second used bigram LMs. Similarly, two PWR systems were developed, one employing a word loop grammar and the second employing bigram LMs. In all cases a multi-layer perceptron neural network was used as a back-end classifier. Various experiments were performed (two-way and four-way language identification). Results indicated that including phonotactic or word LM information increased performance. Furthermore, the word-based systems outperformed the phone-based systems. PWR with LMs gave similar results compared to PWRLM. In a four-way experiment comparing PPR and PWR, both employing trigram LMs, PWR outperformed PPR with a classification accuracy of 84.0% compared to 82.6%.

A PWR approach was similarly followed by Mendoza et al. in the identification of English, Japanese and Spanish [23]. Using approximately 8 hours of training data from 30 speakers for each language, a LVCSR system was developed for each of the three languages. A logistic regression back-end classifier was trained on the scores obtained from the different recognisers using data from the OGI-TS corpus. Evaluation involved identifying 222 10-second utterances and 59 1-minute utterances, evenly distributed among English, Japanese and Spanish. Identification accuracies above 97% were obtained for both durations.

5 Further Advances in Language Identification

Several refinements and improvements of the techniques described in the preceding sections have been developed. As noted in Section 3, Zissman found that the PPRLM outperformed GMM-based language identification systems [5]. The disadvantage of the former is the requirement of orthographically or phonetically transcribed speech data, which is not required for the acoustic modelling approaches. For this reason Torres-Carrasquillo et al. presented several papers dealing

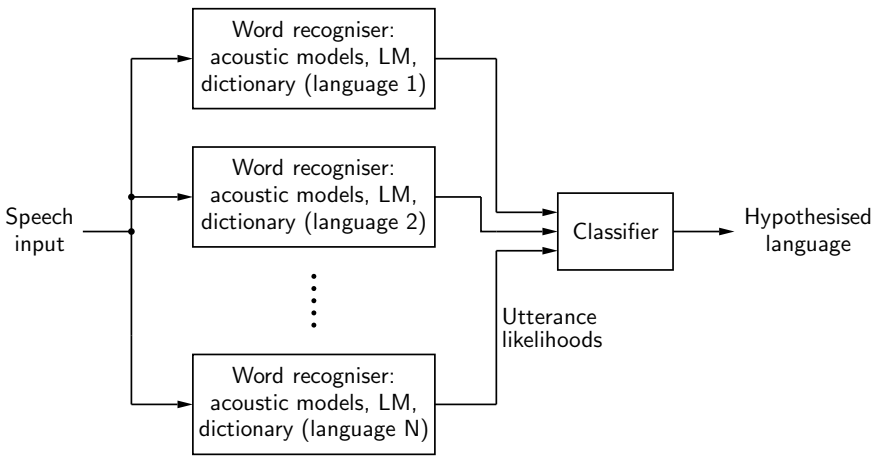

Figure 4: Language identification using parallel word recognition (PWR).

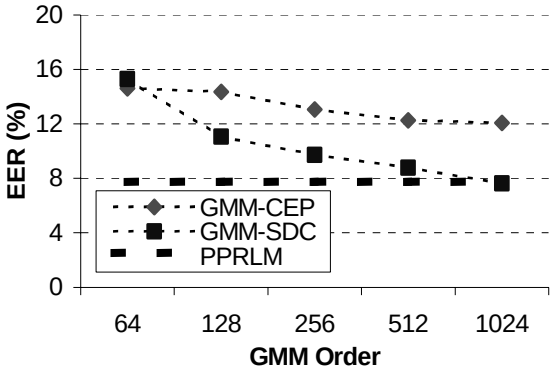

Figure 5: Language identification EERs obtained by Torres-Carrasquillo et al. for GMMs with conventional cepstral features (GMM-CEP), GMMs with SDC features (GMM-SDC) and a PPRLM system on the CallFriend evaluation set. Reproduced from [7].

with improvements and refinements of the acoustic modelling approach.

5.1 GMM Tokenisation and Shifted Delta Cepstra Features

In [24] Torres-Carrasquillo et al. describe a GMM tokenisation approach to language identification. This approach is similar to PRLM (Section 3). The difference is that the front-end phone recogniser is replaced by a GMM tokeniser. Speech is vectorised frame by frame every 10 milliseconds. These frames are then used to train a GMM. During recognition, the output of the front-end recogniser is taken as the index of the Gaussian component with the highest score in the GMM computation for that frame. In analogy to PRLM, these indices are used to train "index language models" for each target language. In [24] several variations of the basic GMM tokenisation approach were evaluated, including the use of a Gaussian back-end classifier which is presented with the different index LM scores, the use of several tokenisers in parallel (in analogy to PPRLM), and the fusion of several models with an additional back-end classifier. Using six GMM tokenisers in parallel, an identification accuracy of 63.7% on the 12 language CallFriend test set was obtained. However, the results obtained using the traditional GMM-based approach gave a slightly higher identification accuracy (64.5%) although a fused system of both approaches gave an accuracy of 73.3%, somewhat comparable to that of a baseline PPRLM system (78.0%). A fusion of the parallel GMM tokenisation approach, the traditional GMM-based approach and PPRLM gave an accuracy of 83.0%.

Many state-of-the-art language identification systems today use shifted delta cepstra (SDC) features which was also evaluated by Torres-Carrasquillo et al. [7]. SDC feature vectors are created by stacking delta cepstra computed across multiple speech frames. The rationale behind this is that temporal information spanning a large number of frames is included in the feature vectors. Four parameters are used to specify the SDC features: the number of spectral coefficients calculated at each time frame N, the time delay and advance for calculating the delta d, the number of blocks for which delta coefficients are concatenated k, and the time delay between consecutive blocks P. The SDC feature vector at time t is given by the concatenation of the vectors $\Delta\mathbf{c}(t + iP)$ for $i = 0, 1, \ldots, k - 1$ where

$$\Delta\mathbf{c}(t) = \mathbf{c}(t + d) - \mathbf{c}(t - d) \tag{4}$$

and $\mathbf{c}(t)$ are the conventional cepstral vectors, resulting in SDC vectors with kN parameters.

Torres-Carrasquillo et al. [7] compared the performance of GMM-based language identification systems using conventional cepstra (with delta-cepstra) and SDC feature vectors. PPRLM using conventional features was also considered. The CallFriend corpus was used and the 12 languages

Table 2: EER (%) on the 1996 NIST LRE evaluation set as obtained in [11].

Approach	30 s	10 s	3 s
1996 PPRLM	9.6	17.8	26.4
2003 PPRLM	5.6	11.9	24.6
GMM	5.1	8.2	16.4
SVM	4.2	11.7	24.0
Fused system	2.7	6.9	17.4

Table 3: EER (%) on the 2003 NIST LRE evaluation set as obtained in [11].

Approach	30 s	10 s	3 s
2003 PPRLM	6.6	14.3	25.5
GMM	4.8	9.8	19.8
SVM	6.1	16.4	28.2
Fused system	2.8	7.8	20.3

included were considered target languages. Results are shown in Figure 5. It is clear that, as the number of mixture components for the GMMs was increased, the EER for the SDC GMMs and the PPRLM approaches became approximately equal (at 1024 mixtures) with the cepstral GMMs performing worst across the different GMM orders considered.

5.2 Further Advancements and NIST LREs

Several other variants, improvements and alternatives to the basic language identification approaches have also been proposed. The development and implementation of these approaches have often come from sites taking part in the more recent NIST LREs. Support vector machine (SVM) classification, for example, is a classification approach that has become part of many of the state-of-the art language identification systems. While more traditional approaches such as GMMs model the PDFs of acoustic features given a specific language, SVMs map inputs into a high dimensional space and then separate classes (languages for our application) with a hyperplane [25]. Discriminative training has also become part of many of the recent language identification systems. The standard maximum likelihood (ML) criterion used to estimate classification models involves maximising the likelihood of the training data given the correct class (language identity or transcriptions in our case). The criterion used for discriminative training, in contrast, is believed to be better related to the identification task itself. For example, discriminative training using the maximum mutual information (MMI) criterion involves maximising the posterior probability of correctly identifying the training utterances [16].

Several authors have successfully employed these techniques. For example, Singer et al. compared PPRLM, GMM-classification with SDC features, SVM classification and fused systems incorporating all these techniques [11]. Evaluations were performed under 1996 and 2003 NIST LRE conditions (see Section 1.3). The authors' 1996 PPRLM system used phone recognisers trained on the OGI-TS corpus. Their 2003 PPRLM system had additional phone models for silences and closures and trigram LMs were added, while their previous system relied on bigram LMs. 2048 mixture GMMs modelling SDC features were trained as well as an SVM classifier. For all systems a Gaussian back-end classifier with LDA normalisation was used. The scores obtained by each of the three approaches were fused using a Gaussian back-end classifier. The three techniques and the fused system were evaluated on both the 1996 and 2003 LRE evaluation sets and the results are shown in Tables 2 and 3.

Table 2 indicates that the 2003 PPRLM system showed a 4% absolute improvement on the 30 second duration test compared to the 1996 PPRLM system. This improvement is the result of using more detailed LMs and additional phone models, i.e. using more specific (detailed, complex) models. The most striking feature of the results in Tables 2 and 3 is the gain obtained using the fused systems incorporating the different knowledge captured by each modelling approach. Similar improvements have also been reported in other papers: Campbell et al. describes a system fusing an SVM system (EER of 6.1% on the 2003 NIST LRE 30 second duration tests) and an SDC GMM-based system (EER of 4.8%) to obtain a fused system yielding an EER of 3.2% [25].

In [16] Burget et al. compared traditional training of GMMs using ML with the discriminative MMI training approach within the 2003 NIST LRE framework. For the 3, 10 and 30 second 2003 LRE test utterances a 2048 mixture ML-trained GMM-based system obtained EERs of 16.3%, 7.9% and 4.7%. The corresponding results for a 128 mixture MMI-trained GMM-based system were 14.8%, 5.5% and 2.1%. Some other approaches were also evaluated in [16], including the use of ergodic HMMs and fusion of the MMI-trained GMM system with a PPRLM system.

Phonotactic modelling approaches have also seen advances. Gauvain et al. proposed that, rather than using the single best hypothesis from the front-end phone recognisers, phone lattices indicating multiple hypothesised phone sequences and associated acoustic scores can be used during both training (for example for training more reliable n-gram LMs) and recognition. In [26] standard PPRLM were compared to PPRLM employing phone lattices. The front-end phone recognisers were trained on 12 hours Egyptian Arabic, 10 hours Spanish and 160 hours American English and LMs were trained on the CallFriend training data. Evaluation was performed in accordance to the 1996 and 2003 NIST LREs. On the 1996 evaluation set, the baseline PPRLM system using the single best hypothesis obtained an EER of 4.9% for 30 second duration test utterances while the PPRLM system employing phone lattices gave 3.2% EER. On the 2003 evaluation set, the baseline PPRLM system gave 6.8% EER while the lattice PPRLM system gave 4.0% EER on the 30 second test utterances. Similar improvements were obtained for the 3 and 10 second test segments. Experiments using a neural network back-end classifier were also considered.

Submissions for the NIST evaluations following the 2003 LRE have grown in complexity and several of the described techniques or variations thereof have been proposed and used by various sites. The performance of systems employing acoustic models trained directly on spectral speech features are often comparable to those of phonotactic modelling approaches. Many sites report considerable gain using fusion techniques incorporating knowledge from both acoustic and phonotactic modelling approaches. For example, Matějka et al. considered and fused ML-trained SDC GMMs, discriminatively trained MMI SDC GMMs, PPRLM, and PPRLM with phone lattices for the 2005 NIST LRE [27]. Torres-Carrasquillo et al. combined several spectral feature and phonotactic identification systems based on discriminatively trained GMMs, SVMs and PPRLM for the 2007 NIST LRE [28] and built on this approach for their submission to the 2009 NIST LRE [29].

5.3 Language Identification of South African Languages

Language identification of four South African languages (Afrikaans, English, Xhosa, Zulu) was considered by Niesler and Willett in [21]. They compared two PPR-based approaches. For the baseline approach HMMs were trained for each language following the standard ML approach. For the second approach language identification-specific discriminatively trained acoustic models were used. The AST databases were used with between 6 and 11 hours of training data for each language and approximately 15 and 25 minutes of development and evaluation data. Test

utterances were short (averaging between 2.0 and 2.6 seconds for each language) compared to durations used for similar studies (for example the 3, 10 and 30 second test segments used in the NIST LREs). The baseline system, using separate phone loop grammars for each language, obtained a language identification accuracy of 79.69% on the evaluation test set. The same system obtained a phone recognition accuracy of 54.15% on the evaluation test set. The best discriminatively trained system showed a language identification accuracy of 79.92% on the evaluation set with a phone recognition accuracy of 54.03%.

6 Dialect and Accent Identification

Although dialect and accent identification have not enjoyed the same amount of attention as the language identification problem, several authors have presented research dealing with the former. Accent identification is often considered more challenging than language identification due to the similar nature of accents [30]. Very similar techniques are however employed for the two problems.

6.1 Phonotactic Modelling

Teixeira et al. performed accent identification experiments for non-native English accents from six European countries using a PPR-based approach [31]. The corpus used was relatively small and consisted of 200 isolated English words repeated two times by 20 speakers from each accent. The training set consisted of 60% of the data while the test set consisted of the remaining 40%. By recognising a test utterance using accent-specific phone HMMs, the authors obtained an average accent identification accuracy of 65.48%. When performing identification followed by accent-specific speech recognition, word recognition accuracies were found to be similar compared to an oracle system where recognition was performed using the accent-specific models corresponding to the accent of the test utterance (a word recognition accuracy of 80.36% compared to the 80.78% accuracy of the oracle system). However, both these approaches were outperformed by a system trained on the pooled data, which gave an accuracy of 85.63%.

In [32], Kumpf and King considered automatic accent identification for three English accents: Australian English, English from first language Lebanese Arabic speakers, and English from first language South Vietnamese speakers. A phone recogniser trained on 2000 sentences of Australian English was used to produce phonetic transcriptions for the training sets which consisted of 1150 Australian English utterances with an average length of 3.2 seconds, 1450 Lebanese Arabic English utterances with an average length of 5.5 seconds, and 1350 South Vietnamese English utterances with an average length of 5.2 seconds. The phonetic transcriptions were used to train phone HMMs as well as phonotactic LMs. These models were employed in a PPR approach to perform accent identification. In an experiment where the identifier had to distinguish between all three accents, an identification accuracy of 76.6% was obtained for 3150 test utterances.

Zissman et al. performed dialect identification for Cuban and Peruvian dialects of Spanish using the basic PRLM configuration [33]. The phone recogniser (employing a phone loop grammar) was trained on the TIMIT corpus (6300 read utterances from 630 speakers) and the Cuban and Peruvian phonotactic LMs were trained on about 3 minutes of speech from approximately 40 Cuban and 20 Peruvian speakers. The test sets similarly consisted of about 3 minutes of speech from approximately 40 Cuban and 20 Peruvian speakers. Using this system, Zissman et al. obtained an average classification accuracy of 84%.

Using PPRLM and a logistic regression back-end classifier, Biadsy et al. [34] performed identification of five dialects of Arabic: Gulf Arabic, Iraqi Arabic, Levantine Arabic, Egyptian Arabic

and Modern Standard Arabic. Phone recognisers for Modern Standard Arabic, English, German, Japanese, Hindi, Mandarin and Spanish were used. Phonotactic LMs were trained from the output of the phone recognisers for each of the dialects for which classification were attempted. The authors achieved an identification accuracy of 81.60% using 30 second test utterances. Training sets consisted of 34.96 hours Gulf Arabic, 18.4 hours Iraqi Arabic, 68.79 hours Levantine Arabic, 47 hours Egyptian Arabic, and 35.54 hours Modern Standard Arabic. Test sets consisted of 6.06 hours Gulf Arabic, 7.33 hours Iraqi Arabic, 10 hours Levantine Arabic, 10 hours Egyptian Arabic, and 12.06 hours Modern Standard Arabic.

6.2 Spectral Feature Modelling

Direct acoustic modelling of the spectral speech features has also been applied to the accent identification problem. Chen et al. used GMMs with MFCC feature vectors for the identification of Beijing, Shanghai, Guangdong and Taiwan accents of Mandarin [35]. Two utterances from 300 speakers in each accent were used for model training. An identification accuracy of 83.2% were obtained using one utterance per test speaker while an accuracy of 86.8% were obtained when using five utterances per speaker. Each utterance was between 3 and 5 seconds long.

In [30], Torres-Carrasquillo et al. performed identification of the same two Spanish dialects considered by Zissman et al. in [33] (see Section 6.1 for a description of [33]). GMMs were trained on SDC feature vectors and were evaluated using the same test set used in [33]. However, an average classification accuracy of only 70% were obtained using this approach, compared to the 84% obtained using the PRLM approach employed in [33].

In [36] Faria describes accent identification followed by speech recognition when distinguishing between native American English and non-native English (speech from Indian, Chinese, Russian, Spanish, German, French and other speakers). A GMM classifier was trained to distinguish between native and non-native speech and based on this distinction the appropriate accent-dependent acoustic models and integrated LM are selected during recognition. Compared to the case where only native acoustic models and a native LM were employed (yielding an average WER of 57.20%), recognition accuracy increased to 53.55% WER when performing identification followed by recognition using the appropriate models. An oracle system gave a WER of 51.70%. The GMM classifier obtained a classification accuracy of 69.5%. Native and non-native training sets both consisted of approximately 60 hours of speech from 374 speakers while test sets consisted of approximately 17 hours of speech from 100 speakers. Faria notes that, in retrospect, accent-independent acoustic and LMs trained on all the data should also have been considered. It might also have been informative if recognition with the native and non-native models running in parallel had been performed.

6.3 Identification of South African English Accents

Although accent identification was not the primary objective in [37], De Wet et al. used a PPR-based approach to automatically classify between three South African English accents: White South African English, English spoken by Nguni (Zulu, Xhosa, Swati, Ndebele) speakers and English spoken by Sotho (Northern Sotho, Southern Sotho, Tswana) speakers. The focus of that research was to determine if accented speech from Nguni and Sotho speakers should be treated as a single variety when developing ASR systems, or whether they should be treated separately. It was found that although the automatic accent identification system could distinguish between White South African English and Black South African English (accented English speech from either Nguni or Sotho speakers), the systems could not distinguish between Nguni and Sotho

varieties of Black South African English. The conclusion was drawn that English from mother-tongue Nguni and Sotho speakers should be considered a single variety when developing ASR technology for these accents. A subset of the AST databases was used with approximately 2.6 hours of speech for training for each of the three accents and approximately 13 minutes of test data per accent.

Using the AST databases, Du Toit considered transcription-less accent identification for the five South African English accents: Afrikaans English, Black South African English, Cape Flats English, White South African English, and Indian South African English [1]. He considered various configurations utilising GMMs and ergodic HMMs. Except for first-order ergodic HMMs, higher order HMMs were also considered. Evaluation was performed on seven test durations (2, 4, 10, 30, 60, 120, and 300 seconds). Shorter utterances were concatenated in order to obtain data for all these durations. The best performing system employed second-order ergodic HMMs with full covariance Gaussian observation PDFs and showed identification accuracies of 52.8%, 58.92%, 70.81%, 86.13%, 94.50%, 98.16%, 98.67% for the 2, 4, 10, 30, 60, 120, and 300 second test segments respectively. This study was conducted before the orthographic and phonetic transcriptions for the AST databases were completed, necessitating the direct modelling of spectral features.

7 Summary and Conclusions

In this document we gave a brief overview of literature dealing with language, dialect and accent identification. Approaches to language identification can be roughly divided into two groups: acoustic modelling where spectral features of different languages are modelled directly, and phonotactic modelling where speech is tokenised into phone strings and scored using different language models (LMs). Early research showed that acoustic models such as Gaussian mixture models (GMMs) using conventional cepstral features were outperformed by phonotactic modelling approaches such as phone recognition followed by language modelling (PRLM), parallel phone recognition followed by language modelling (PPRLM), and parallel phone recognition (PPR) [33]. However, Torres-Carrasquillo et al. showed that GMMs using shifted delta cepstra (SDC) features can result in comparable performance to that of PPRLM without the requirement of transcribed speech data [24].

Driven by the more recent NIST language recognition evaluations (LREs), several variants, improvements and alternatives to the basic language identification techniques have been proposed. Support vector machines (SVMs), discriminative training, and more complex back-end classifiers are part of most of the state-of-the-art language identification systems. By performing back-end classification on the scores obtained from different models, a fused system can take advantage of the different knowledge captured by several modelling approaches. Using such fused systems, several authors have reported significant gain compared to isolated modelling approaches [11, 25, 27, 28, 29].

Many of the techniques developed for language identification are directly relevant to the dialect and accent identification problems. For example, Zissman et al. performed dialect identification for Cuban and Peruvian dialects of Spanish using the basic PRLM configuration [33]. In [30], Torres-Carrasquillo et al. performed identification of the same two Spanish dialects using SDC-based GMMs but lower identification accuracies were reported compared to the PRLM approach. Several other authors have also applied direct spectral feature modelling [35, 36] and phonotactic modelling [31, 32, 34] to the dialect and accent identification problems.

For a specific language, dialect or accent identification problem, it seems as if the most appropriate isolated modelling technique is dictated by the availability of data. Where no transcribed

speech data are available direct acoustic modelling approaches such as SDC-based GMMs seems most appropriate. If transcribed data for languages different from the target languages are available PRLM or PPRLM can be implemented. Where transcribed speech data for the target language are available PPR or parallel word recognition (PWR) could be considered.

Table 4: Summary of literature dealing with language, dialect and accent identification.

Authors	Languages/Dialects/Accents	Approach(es)	Corpora	Findings/Conclusions
Zissman [17]	Various identification problems were considered, including: identification of Mandarin, Tamil and Japanese; identification of Russian, German and Mandarin; classification of 20 languages in the CCITT corpus.	GMMs, ergodic HMMs.	For Mandarin, Tamil and Japanese each 50 to 60 minutes conversational speech; Russian, German, Mandarin each 15 to 20 messages; 20 CCITT languages each 16 utterances about 8 seconds long.	Ergodic HMMs performed similar to GMMs, indicating that the temporal modelling capabilities of the HMMs were not exploited.
Zissman [5]	First identification problem: English, Japanese and Spanish. Second identification problem: 10 languages (OGI-TS languages excluding Hindi).	GMMs, PRLM, PPRLM, PPR.	OGI-TS training partition for training and development partition for testing. Also used OGI-TS phonetic transcriptions.	PPR and PPRLM gave similar performance followed by PRLM. GMMs performed worst.
Wong and Siu [18]	English, German, Hindi, Japanese, Mandarin, Spanish.	PRLM using a English phone recogniser.	OGI-TS corpus.	57% and 78% classification accuracies for 10 and 45 second utterances.
Yan and Barnard [19]	English, German, Hindi, Japanese, Mandarin, Spanish.	PPRLM with a neural network back-end classifier.	OGI-TS corpus, specifically the phonetic transcriptions.	81% and 92% classification accuracies for 10 and 45 second utterances.
Fernández et al. [20]	English, Spanish.	PPRLM using English and Spanish phone recognisers.	Phone recogniser training: 4.7 h English, 7.1 h Spanish. LM training: 1 h English, 0.9 h Spanish. Test set: 0.9 h English, 0.9 h Spanish.	Classification accuracy of 92.6%.
Schultz et al. [22]	English, German, Japanese, Spanish.	PPR without LM (i.e. a phone loop), PPR with LM, PWR without LM, PWR with LM, PWRLM.	6.9 h (7644 utterances) English, 30.5 h (12 292 utterances) German, 8 h (3311 utterances) Japanese, 10.7 h (5730 utterances) Spanish.	Better performance with LMs. Better performance for word-based identification (PWR, PWRLM). Using trigram LMs, 82.6% and 84.0% accuracies using PPR and PWR.
Mendoza et al. [23]	English, Japanese, Spanish.	LVCSR-based identification (PWR). Used a logistic regression back-end classifier.	Training: 8 h from 30 speakers in each language. Back-end classifier training: OGI-TS data. Test set: 222 10 s, 59 1 min segments in total from OGI-TS.	Classification accuracies higher then 97% for both 10 second and 1 minute segments.
Torres-Carrasquillo et al. [24]	The 12 CallFriend languages.	GMM tokenisation with Gaussian back-end classifier.	GMM tokenisers and LMs trained on training partition of CallFriend. Back-end classifier trained on development set. CallFriend evaluation set used for evaluation.	Accuracy of 63.7% with GMM tokenisation; 64.5% with GMM acoustics; 73.3% with fusion of preceding two; 83.0% with fusion of tokenisation, acoustic, PPRLM.

Authors	Languages/Dialects/Accents	Approach(es)	Corpora	Findings/Conclusions
Torres-Carrasquillo et al. [7]	The 12 CallFriend languages.	Compared GMMs with conventional cepstral features, GMMs using SDC features, PPRLM. Also compared GMM tokenisation with cepstral and SDC features. All systems had Gaussian back-end classifiers with LDA normalisation.	CallFriend training set for GMMs and LMs, CallFriend development set for back-end classifiers. Unclear what was used for phone recognisers in PPRLM, possibly OGI-TS corpus.	For GMMs with 1024 mixtures, PPRLM and GMMs with SDC features both gave 8% EER while GMMs with cepstral features gave 12% EER. GMM tokenisation performed better using SDC compared to cepstral features.
Singer et al. [11]	The 12 CallFriend languages. Tested under 1996 and 2003 NIST LRE conditions.	PPRLM, SDC GMM classification, SVM classification and a fused system. All systems had Gaussian back-end classifiers with LDA.	Phone recognisers were trained on OGI-TS corpus, LMs from the CallFriend training set, and back-end classifiers on the 1996 and 2003 NIST LRE development data. Evaluation performed on corresponding NIST test sets.	EERs of 5.6%, 5.1%, 4.2%, 2.7% for PPRLM, GMM, SVM and fused systems on 30 second 1996 NIST data. EERs of 6.6%, 4.8%, 6.1%, 2.8% for PPRLM, GMM, SVM and fused systems on 30 second 2003 NIST data.
Campbell et al. [25]	The 12 CallFriend languages. Tested under 2003 NIST LRE conditions.	SDC GMM classification, SVM classification and a fused system.	CallFriend training set and the 1996 and 2003 NIST LRE development data.	EERs of 6.1%, 4.8%, 3.2% for SVM, GMM and fused systems on 30 second 2003 NIST data.
Burget et al. [16]	The 12 CallFriend languages. Tested under 2003 NIST LRE conditions.	Compared ML and discriminatively trained (MMI) GMMs, and some other variations.	CallFriend training set. Evaluated on the 2003 NIST LRE evaluation set.	EER of 4.7% for 2048 mixture ML-trained GMMs and EER of 2.1% for 128 mixture MMI-trained GMMs on 30 second test set.
Gauvain et al. [26]	The 12 CallFriend languages. Tested under 1996 and 2003 NIST LRE conditions.	Compared standard PPRLM and PPRLM employing phone lattices instead of the single best hypothesis from the phone recognisers.	Phone recognisers trained on 12 h Egyptian Arabic, 10 h Spanish, 160 h American English. CallFriend training set for LMs. 1996 and 2003 NIST LRE evaluation sets.	1996 NIST evaluation: 4.9% EER for PPRLM vs. 3.2% for PPRLM using phone lattices. 2003 NIST evaluation: 6.8% EER for PPRLM vs. 4.0% for PPRLM using phone lattices for 30 second test utterances.
Niesler and Willett [21]	Afrikaans, English, Xhosa, Zulu.	A PPR-based approach comparing ML-trained acoustic models (HMMs) with language identification-specific discriminatively trained acoustic models.	AST databases with 6 to 11 hours for each language for training and approximately 15 and 25 minutes per language for development and evaluation respectively.	Small improvement for language identification accuracy using discriminatively trained models (76.69% vs. 79.92%).

Authors	Languages/Dialects/Accents	Approach(es)	Corpora	Findings/Conclusions
Teixeira et al. [31]	British, Danish, German, Italian, Portuguese and Spanish accents of English.	PPR. Also considered accent identification followed by word recognition.	200 isolated English words repeated two times by 20 speakers from each accent. Training set: 60% of the data. Test set: 40%.	Identification accuracy of 65.48%. Identification followed by recognition performed similar to oracle system. A pooled system performed best.
Kumpf and King [32]	Australian English, English from first language Lebanese Arabic (LA) speakers, and English from first language South Vietnamese (SV) speakers.	PPR. Used an Australian English phone recogniser to obtain phonetic transcriptions for each database, from which accent-specific phone HMMs and LMs were trained.	1150 Australian English utterances (3.2 s average), 1450 LA utterances (5.5 s), 1350 SV utterances (5.2 s). 2000 sentences used to train initial phone recogniser. 3150 test utterance were used for evaluation.	Identification accuracy of 76.6%.
Zissman et al. [33]	Cuban and Peruvian dialects of Spanish.	PRLM using a English phone recogniser trained on the TIMIT corpus.	Phonotactic LMs trained on about 3 minutes of speech from 40 Cuban and 20 Peruvian speakers. Test set similar size to training set.	Identification accuracy of 84%.
Biadsy et al. [34]	Five dialects of Arabic: Gulf Arabic, Iraqi Arabic, Levantine Arabic, Egyptian Arabic and Modern Standard Arabic.	PPRLM, using English, German, Hindi, Japanese, Mandarin and Spanish phone recognisers, and a logistic regression back-end classifier.	Phone recognisers trained on OGI-TS corpus. LMs trained on between 20 and 70 hours from each Arabic dialect. Test set: between 6 and 12 hours from each dialect.	Identification accuracy of 81.60% using 30 second test utterances.
Torres-Carrasquillo et al. [30]	Cuban and Peruvian dialects of Spanish.	GMMs with SDC feature vectors.	The same corpora as in [33].	Identification accuracy of only 70%, compared to the 84% obtained using PRLM in [33].
Chen et al. [35]	Beijing, Shanghai, Guangdong and Taiwan accents of Mandarin.	GMM-based classification using MFCC feature vectors.	Training set: two utterances from 300 speakers in each accent. Test set: 50 utterances from 60 speakers in each accent. Each utterance was 3 to 5 seconds.	83.2% identification accuracy for one utterance per test speaker. 86.8% identification accuracy for five utterances per test speaker.
Faria [36]	Native American English and non-native English.	Used a GMM classifier followed with accent-dependent speech recognition.	Native and non-native training sets: 60 hours of speech from 374 speakers. Test sets: 17 hours from 100 speakers.	Identification accuracy: 69.5%. Word recognition using oracle system: 51.70% WER. Identification with recognition: 53.55% WER. Native models: 57.20% WER.
De Wet et al. [37]	Three South African English accents: White South African English, English spoken by Nguni speakers and English spoken by Sotho speakers.	PPR.	A subset of the AST databases. Training: 2.6 hours per accent. Test set: 13 minutes per accent.	Identification accuracy of 62.2%. Concludes that English from Nguni and Sotho speakers should be considered a single variety.
Du Toit [1]	The five South African English accents.	GMMs; first-, second- and third-order ergodic HMMs.	The AST databases.	Second-order ergodic HMM with full covariance Gaussian PDF gave best results: 86.13% identification accuracy on 30 second test utterances.

References

[1] A. du Toit, "Automatic classification of spoken South African English variants using a transcription-less speech recognition approach," Master's thesis, Stellenbosch University, 2003.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97)*, Rhodes, Greece, 1997, pp. 1895–1898.

[3] A. F. Martin and A. N. Le, "The current state of language recognition: NIST 2005 evaluation results," in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey '06)*, San Juan, Puerto Rico, 2006, pp. 1–6.

[4] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '92)*, Alberta, Canada, 1992, pp. 895–898.

[5] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[6] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, pp. 115–124, 2001.

[7] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, CO, 2002, pp. 89–92.

[8] Linguistic Data Consortium, "Linguistic Data Consortium catalog," 2011. [Online]. Available: http://www.ldc.upenn.edu/Catalog

[9] A. F. Martin and M. A. Przybocki, "NIST 2003 language recognition evaluation," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2003)*, Geneva, Switzerland, 2003, pp. 1341–1344.

[10] National Institute of Science and Technology (NIST), "The 1996 language recognition evaluation plan," 1996.

[11] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2003)*, Geneva, Switzerland, 2003, pp. 1345–1348.

[12] National Institute of Science and Technology (NIST), "The 2003 language recognition evaluation plan," 2003.

[13] ——, "The 2005 language recognition evaluation plan," 2005.

[14] ——, "The 2007 language recognition evaluation plan," 2007.

[15] ——, "The 2009 language recognition evaluation plan," 2009.

[16] L. Burget, P. Matějka, and J. Černocký, "Discriminative training techniques for acoustic language identification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, Toulouse, France, 2006, pp. 209–212.

[17] M. A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93)*, vol. 2, Minneapolis, MN, 1993, pp. 399–402.

[18] K. Wong and M. Siu, "Automatic language identification using discrete hidden Markov model," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2004 – ICSLP)*, Jeju Island, Korea, 2004, pp. 1633–1636.

[19] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, Detroit, MI, 1995, pp. 3511–3514.

[20] F. Fernández, R. de Córdoba, J. Ferreiros, V. Sama, L. F. D'Haro, and J. Macìas-Guarasa, "Language identification techniques based on full recognition in an air traffic control task," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2004 – ICSLP)*, Jeju Island, Korea, 2004, pp. 1565–1568.

[21] T. R. Niesler and D. Willett, "Language identification and multilingual speech recognition using discriminatively trained acoustic models," in *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing (MULTILING 2006)*, Stellenbosch, South Africa, 2006.

[22] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based language identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, Atlanta, GA, 1996, pp. 781–784.

[23] S. Mendoza, L. Gillick, Y. Ito, S. Lowe, and M. Newman, "Automatic language identification using large vocabulary continuous speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, Atlanta, GA, 1996, pp. 785–788.

[24] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller, "Language identification using Gaussian mixture model tokenization," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, Orlando, FL, 2002, pp. 757–760.

[25] W. M. Campbell, J. P. Campbell, R. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[26] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2004 – ICSLP)*, Jeju Island, Korea, 2004, pp. 25–28.

[27] P. Matějka, L. Burget, P. Schwarz, and J. Černocký, "Brno university of technology system for NIST 2005 language recognition evaluation," in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey '06)*, San Juan, Puerto Rico, 2006, pp. 1–7.

[28] P. A. Torres-Carrasquillo, E. Singer, W. M. Campbell, T. P. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. Sturim, "The MITLL NIST LRE 2007 language recognition system," in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 719–722.

[29] P. A. Torres-Carrasquillo, E. Singer, T. P. Gleason, A. McCree, D. A. Reynolds, F. Richardson, and D. Sturim, "2009 language recognition system," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, Dallas, TX, 2010, pp. 4994–4997.

[30] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey '04)*, Toledo, Spain, 2004, pp. 41–44.

[31] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, PA, 1996, pp. 1784–1787.

[32] K. Kumpf and R. W. King, "Automatic accent classification of foreign accented Australian English speech," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, PA, 1996, pp. 1740–1743.

[33] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, Atlanta, GA, 1996, pp. 777–780.

[34] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages (CASL '09)*, Athens, Greece, 2009, pp. 53–61.

[35] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *Workshop on Automatic Speech Recognition and Understanding (ASRU '01)*, Madonna di Campiglio, Italy, 2001, pp. 343–346.

[36] A. Faria, "Accent classification for speech recognition," in *Proceedings of the Second Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI '05)*, Eddinburgh, UK, 2005.

[37] F. De Wet, P. Louw, and T. R. Niesler, "Human and automatic accent identification of Nguni and Sotho Black South African English," *South African Journal of Science*, vol. 103, no. 3/4, pp. 159–164, 2007.