Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 20 (2006) 276–302
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A syllable-scale framework for language identification

Terrence Martin *, Brendan Baker, Eddie Wong, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, 2 George Street, GPO Box 2434, Brisbane, Qld 4001, Australia

Received 1 November 2004; received in revised form 26 July 2005; accepted 29 July 2005
Available online 31 August 2005

Abstract

Whilst several examples of segment based approaches to language identification (LID) have been published, they have been typically conducted using only a small number of languages, or varying feature sets, thus making it difficult to determine how the segment length influences the accuracy of LID systems. In this study, phone-triplets are used as crude approximates for a syllable-length sub-word segmental unit. The proposed *pseudo-syllabic* length framework is subsequently used for both qualitative and quantitative examination of the contributions made by acoustic, phonotactic and prosodic information sources, and trialled in accordance with the NIST 1996 LID protocol. Firstly, a series of experimental comparisons are conducted which examine the utility of using segmental units for modelling short term acoustic features. These include comparisons between language specific Gaussian mixture models (GMMs), language specific GMMs for each segmental unit, and finally language specific hidden Markov models (HMM) for each segment, undertaken in an attempt to better model the temporal evolution of acoustic features. In a second tier of experiments, the contribution of both broad and fine class phonotactic information, when considered over an extended time frame, is contrasted with an implementation of the currently popular parallel phone recognition language modelling (PPRLM) technique. Results indicate that this information can be used to complement existing PPRLM systems to obtain improved performance. The pseudo-syllabic framework is also used to model prosodic dynamics and compared to an implemented version of a recently published system, achieving comparable levels of performance.

* Corresponding author. Tel.: +61 7 38641414.
E-mail addresses: tl.martin@qut.edu.au (T. Martin), bj.baker@qut.edu.au (B. Baker), ee.wong@qut.edu.au (E. Wong), s.sridharan@qut.edu.au (S. Sridharan).

0885-2308/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2005.07.004

1. Introduction

The goal of automatic language identification (LID) is to identify a language based on a sample of speech. There are many potential applications for an accurate LID system. Aside from interest by various governmental agencies focused on meeting specific needs for surveillance and national security, LID systems can also be used to assist telecommunication companies in handling foreign language calls. For instance, these systems can be used to detect the spoken language prior to redirecting callers to an appropriate translator or machine translation system. An example quoted in Muthusamy (1993) also illustrates that LID systems can help to save lives. LID systems can be used to quickly redirect the caller to an appropriate translator, thus minimising potentially life threatening delays which can occur when the emergency operator cannot understand the language spoken by the caller.

In order to produce accurate LID systems, researchers have sought to exploit some of the cues that humans use to discriminate between languages. The techniques employed have generally been based on exploiting some derivation of acoustic phonetics, phonotactics and prosodic information.

The use of acoustic phonetics for LID is based on exploiting spectral differences in the realisation of phonemes between languages. Phonotactic information specific to a language is contained within the statistical dependencies inherent in phonetic chains (Navratil, 2001), and accounts for the probability distribution of phonetic elements as well as any constraints imposed upon sequences of phonemes. Finally, prosodic information consists of the relative durations and amplitudes of sonorant segments, their spacing in time and patterns of pitch change within and across segmental units (Muthusamy, 1993).

The research outlined in this paper examines the contributions of these information sources under a unified probabilistic framework, using a segment based approach. This approach is based heavily on work proposed by Hazen and Zue (1997), albeit with some differences in implementation, and examined using more recently available databases.

In Hazen and Zue (1997), the segmental unit used was the phone, however in the research outlined in this paper, an extended segmental unit is used, based on a phone-triplet of *broad phonetic* classes. This temporal length was chosen in an attempt to crudely approximate the length of the syllable. The selection of broad class phones was motivated by the need to provide a balance between the competing requirements of model set size and the practical limitations associated with obtaining adequate training data and accurate segmentation.

Whilst the use of broad class phones and segment based approaches to LID have been well researched, a majority of the work has been trained and evaluated using smaller databases, or with varying sets of languages. In Section 2, a motivation for revisiting this approach is provided against a backdrop of research previously conducted in the field of LID. Section 3 provides an outline of the data used for training and testing both the recognition and identification systems.

Sections 4 and 5 then provide a detailed description of the phone-triplet segmentation process. This framework is then used to conduct both qualitative and quantitative evaluation of the contributions made by acoustic, phonotactic and prosodic information sources, and is trialled in accordance with the NIST 1996 LID protocol.

A number of techniques have been proposed for improving the acoustic modelling in LID systems, however once again reported results have been difficult to compare due to differences

in languages used and implementation variations. Accordingly, a series of experimental comparisons were conducted which examined the utility of using segmental units for modelling short term acoustic features using both GMM and HMM modelling techniques. In order to examine whether the proposed extended segmental unit is any better than a phone-length segmental unit, an appropriate comparison is also provided. Section 6 presents details of these experiments.

Section 7 outlines experiments based on simple statistical modelling of the frequency of occurrence of the syllabic events, and how it can be used to complement existing PPRLM approaches. Once again, utilising the syllable-length framework, Section 8 presents results for experiments using a pitch and energy trajectory feature set and a comparison with an implemented version of the technique presented by Adami and Hermansky (2003).

Finally, in Section 9, a series of fusion experiments are outlined using an appropriate selection of the baseline systems presented in Sections 6–8. Conclusions which arise from the presented experiments are then provided in Section 10.

2. Background and motivation

The techniques employed to improve LID performance have generally been based on exploiting some derivation of acoustics, phonotactics and prosodic information. In the standardised set of LID evaluations conducted by NIST in 2003, most systems incorporated some form of short-term acoustic modelling, typically using language dependant Gaussian mixture models (GMMs) adapted from universal background models (UBMs) (Reynolds, 1997), as well as phonotactic modelling such as phone recognition language modelling (PRLM) (Tucker et al., 1994; Hazen and Zue, 1997) or the more elaborate parallel phone recognition language modelling (PPRLM) (Zissman and Singer, 1994; Yan et al., 1996). Whilst systems incorporating the use of prosodics have not been predominant in NIST evaluations, there has been renewed interest in their utility, with an interesting paper based on modelling prosodic dynamics presented recently by Adami and Hermansky (2003).

More recently researchers are looking to utilise higher level knowledge, in an attempt to exploit the complementary nature of different feature sets, in the hope it may lead to more robust identification systems. Segment based approaches offer a convenient means for examining the individual contributions under a common framework, allowing qualitative and linguistically meaningful comparisons. There are a number of candidates available for use as a segmental unit, with the phone, syllable and word the most obvious.

A number of segment based approaches have been proposed in LID literature, and a well reasoned mathematical formulation, based around a segment based approach to LID, is presented in (Hazen and Zue, 1997). In Hazen's work, information sources were broken into two categories; segmental or prosodic. Segmental information was subsequently subdivided into contributions from acoustic phonetics or phonotactics. Hazen selected the phone as the segmental unit. In contrast, the work presented in this paper is based on an extended segmental unit, roughly based around the syllable. A number of reasons underpinned the selection of this extended unit. One motivation stemmed from a desire to recover more information from short term acoustic features and recently published research by Torres-Carrasquillo et al. (2002) indicates that this may be more effectively obtained by using an extended time-frame.

In that study, the use of the shifted delta cepstrum (SDC) technique overcame the inability of GMM modelling to capture temporal dependencies. Torres-Carrasquillo et al. (2002) concatenated sequences of spectral features to form an extended feature vector, in effect providing a means for modelling temporal patterns. This implicit segmentation via the feature set led to state of the art performance. In a subsequent study using SDC (Singer et al., 2003), it was stated that in-house experimentation conducted by a third party resulted in the selection of a feature vector which spanned seven 30 ms frames. This coincides approximately with the length of a syllable, and suggests that there is at least good circumstantial evidence to support the concept of using an extended unit. Whilst the results achieved by SDC are impressive, it would be useful to be able to associate those temporal sequences which contain the most discriminatory information, and a linguistically meaningful label. In an attempt to achieve this, *broad* phone-triplets were selected as an extended segmental unit, based crudely around representing a common syllable-length and conveniently coinciding with the approximate length of the SDC based feature vectors.

The decision to use *broad* phonetic classes in preference to *fine* phonetic classes was based mainly according to practical considerations. It was decided to arbitrarily fix the length of the syllable to three phones, rather than to produce true syllabic segments. Whilst this does not necessarily reflect the true syllabic structures and lengths which exist across languages, the expertise required and difficulties involved in performing multilingual syllabification necessitated this simplification.

Given this arbitrary three phone length, the degree of phonetic resolution used for each phone required determination. Unfortunately, as the size of the phone-set increases, the possible number of triplets expands exponentially. Additionally, the accuracy of phone recognition systems generally degrade as the phone model inventory increases in size. Accordingly, the phone-set was restricted to four broad classes. It is possible that this crude clustering may confuse both the model space and the usefulness of the temporal information, however previous research has revealed that even at the broad class level, quite successful discrimination can be achieved.

For example, in a landmark study, House and Neuberg (1977) examined the constraining power of low-level phonotactic information by using an ergodic HMM to model *text derived* sequences of broad phonetic categories for each language. Using just five broad classes on an eight language identification task they achieved perfect LID. This study utilised manually generated transcriptions, and hence assumed an ideal condition of perfect transcription accuracy. Regardless of this shortcoming, this study was an important concept demonstration, and highlighted that quite powerful discriminative information exists even at the broad phone level. Subsequent *real world* extensions to this idea were conducted by Li and Edwards (1980), Hieronymus and Kadambe (1996) and Yan and Barnard (1995) by using speech recognition systems to derive speech transcriptions. Understandably the language identification results degraded, but the level of performance achieved provided additional validation for the idea that sequences of broad phonetic categories possess discriminatory information.

A comprehensive examination of the contributions that broad class phonetic events make to the LID task was subsequently undertaken in the Ph.D. thesis work in Muthusamy (1993). Muthusamy noted that languages can be described using phrases such as guttural, nasal, singsong or rhythmic. Based on the principal that any segmental unit which seeks to incorporate these characteristics needs to span multiple phonemic events, but not necessarily contain a fine degree of phonetic resolution, Muthusamy examined various features and segmental units. Aside from

features based around broad phonetic classes, he also examined segmental units based on extended time frames including phone-pairs and phone-triplets. Muthusamy found that phone-pairs produced the best performance. Whilst Muthusamy concluded that broad class phonetic events do contain language discriminatory information, he suggested that better performance may be possible if the degree of phonetic resolution was expanded to include finer phonetic classes.

Berkling however suggested that the contribution of both broad and fine phonetic features were not necessarily mutually exclusive (Berkling, 1996). She commented that both the length of sequence information and the degree of phonetic resolution required for LID can vary, according to the languages under consideration, and to illustrate her point, Berkling outlined several language-pair analogies. English and German, for example, are similar in that they have similar consonant-clusters and vowel frequencies. In this case, a more refined phonetic representation is required to discriminate between these languages. In contrast, Japanese and Chinese have a highly constrained syllabic structure and may be discriminated quite effectively using a much broader phonetic representation. Finally, Berkling also commented that some languages can be discriminated by analysing short sequences, while other languages require analysis over an extended period. A further extension to this idea is that the sequences of phonetic classes which are important when using prosodic information may be quite different to what is required using phonotactic or acoustic information sources.

It is possible to examine whether different sources of complementary information are contained within crude and fine phonetic detail; over both short and long time frames, for *any* of the three information sources. However, for the research reported in this paper, this concept was only explored using phonotactic information and is outlined in Sections 7 and 9. However, by using the syllabic framework, it was possible to examine the discriminative power provided by each of the syllables, and whether this discriminative power is the same for acoustic, phonotactic and prosodic sources. These differences have been highlighted in Sections 6 and 8.

As mentioned earlier, one of the motivations for this research was to try and improve the amount of information extracted from short-term acoustic features. One of the downfalls of the commonly used GMM/UBM technique is that it cannot model temporal dependencies between frames. While both GMMs and HMMs both make the assumption of frame independence, Hidden Markov modelling explicitly incorporates piece-wise sequence information via the left to right state topology in combination with state transition statistics, and may be better suited for modelling this trajectory information. Of course, the use of SDC provide a means for circumventing this problem, however we have chosen to re-examine the use of HMM's for capturing this information.

Several researchers have used hidden Markov models (HMMs) in an attempt to overcome the shortfall of GMM modelling. For instance in Zissman and Singer (1995), a comparison was conducted on the performance of ergodic HMMs versus GMM modelling, with only one HMM and GMM built per language. Zissman and Singer (1995) found that the sequential modelling capability of HMMs did not realise any improvements over GMM modelling. In contrast, Nakagawa et al. (1994) found that the use of HMMs gained improved performance in comparison to GMMs. Notably, the HMMs produced by Zissman for each language modelled the state observation probabilities using the same GMM, with the only difference between states existing in the allocation of mixture weights. As noted by Zissman, this tied-state topology was necessary to limit the number of parameters requiring estimation due to the size of database used for training. However,

with the release of the CallFriend database (Linguistic Data Consortium, 1996), more data are available and hence it is possible to use improved HMM topologies.

For those studies that also incorporated segment based approaches for modelling short-term acoustics, the phone has generally been chosen as segmental unit. For example, in research reported by Hieronymus, single state ergodic HMMs were produced for each phone. However, the features modelled were only based on the statistical parameters of duration, such as mean, variance, maximum and minimum values. Thus it is difficult to compare the results achieved in this study with the currently preferred method of using GMMs for modelling cepstral-based features.

In a more elaborate implementation, Parris and Carey (1995) modelled Mel-scale frequency cepstral coefficients (MFCCs) by producing HMMs for context-independent monophones for the Dutch, English and Norwegian languages. These models were then used to decode the speech and a maximum likelihood approach used to determine the most likely language. Finally in Mendoza et al. (1996), context-dependent triphone models were produced and evaluated using a similar technique on the English, Japanese and Spanish languages. The results achieved in these studies indicate that pre-segmenting speech prior to modelling can provide discriminative advantages, and the use of HMMs to model the temporal evolution of these features is viable. However, it is difficult to determine whether using HMMs can provide any benefit over GMM modelling of the acoustics as the studies which did report direct comparisons were conflicting. Additionally, many of the reported results were evaluated using a limited set of languages or much smaller amount of training data. Accordingly in Section 6, this approach has been revisited. Comparisons have been conducted between the typical implementation of a GMM/UBM system (as used by this research laboratory in the NIST 2003 evaluation), a system which produces language specific GMM's for each phone-triplet, and finally a system using HMM's to model each phone-triplet.

Most of the previous paragraphs have highlighted the research contributions made towards improving the modelling of short-term acoustics. However, an equally important body of work has focused on phonotactic modelling. Importantly, the use of *n*-gram modelling to capture information contained in the distribution of individual phonemes and any language specific contextual constraints is an important and powerful technique. Systems have been proposed which use single-language phone recognition systems followed by phonotactic modelling in Tucker et al. (1994), or single recognition system based on a language independent phone set (Hazen and Zue, 1997). Zissman and Singer (1994) and Yan et al. (1996) both extended this idea by using a series of single language recognisers in parallel (PPRLM) and the degree of success achieved has seen this technique implemented by a large number of entrants to the NIST LID evaluations. Accordingly, an implementation of the PPRLM system used in the NIST 2003 LID evaluation by QUT, as outlined in Wong and Sridharan (2003), is used as a baseline for comparison purposes. In an attempt to examine the ideas proposed by Berkling (1996), as outlined earlier, a series of additional experiments were conducted using unigram statistics for phone-triplets, at both the broad class and fine phonetic level. This is outlined in Section 7.

Finally, a smaller body of research exists which examines the discriminatory contribution of prosodic information, but work conducted by Foil (1986), Goodman et al. (1989), Savic et al. (1991), Hutchins and Thyme-Gobbel (1994) all provide an indication of its utility. In more recent work, Adami and Hermansky (2003) combined prosodic information such as pitch, energy and articulatory event duration with phonotactics, achieving some interesting results. Aside from

providing a simple, yet effective technique for capturing information about prosodic dynamics, Adami's work reinforced previous studies by House and Neuberg (1977), Muthusamy (1993) and Berkling (1996) which indicate that discriminative information is contained even at the broad phonetic level. Whilst these studies have all illustrated that useful discriminative information can be gleaned by prosody based systems, once again it is difficult to analyse how much these systems contribute in comparison to other sources of information, especially when examined under the more recent and expansive LID protocols. Accordingly in Section 8, using the same common syllabic framework, a system which modelled pitch dynamics was built and subsequently compared to a system implemented in accordance with the basic principles outlined by Adami. This system is also subsequently used in a series of fusion experiments in both Sections 8 and 9.

3. Database and performance evaluation

Two main data sources were utilised in this study: the OGI multi-language telephone speech (OGI-MLTS) corpus and CallFriend. The OGI-MLTS corpus contains phonetic transcriptions for six languages (Muthusamy et al., 1992); English, Hindi, Spanish, Mandarin, German and Japanese. This corpus was used for both preliminary idea validation and subsequently used as training data for the core multi-lingual broad-phone recogniser.

Table 1 provides information detailing the distribution of data for training, development and evaluation. The data from the original distribution did not have any Hindi utterances, but a subsequent extended data release included supplementary utterances for Hindi, English, Spanish and German. Accordingly, 13 Hindi utterances from the extended data set were combined with the original development and test data to produce new sets. All other extended data were added to the training data.

The majority of LID experiments carried out in this study were conducted on an additional corpus, the LDC CallFriend telephone speech database (Linguistic Data Consortium, 1996). Experiments were carried out according to protocols defined for the NIST 1996 LID evaluation. The CallFriend database consists of 12 different languages (Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese) of which three have a second dialect (English, Mandarin and Spanish). For this study, when training data were available for two dialects, the training data were merged, resulting in double the amount of training data for these languages. The evaluation tests the performance of the LID system on utterances of duration 3, 10 and 30 s, with the number of test segments per duration being 1503, 1502 and 1492, respectively. The development set had 1174, 1172 and 1174, respectively.

Table 1
Number of utterance for phonetically transcribed OGI training, development and evaluation sets

	Language					
	English	German	Hindi	Japanese	Mandarin	Spanish
Train	110	62	55	46	45	72
Development	18	20	0	2	12	17
Test	19	21	13	17	13	17

Comparisons of language identification performance in this study are achieved through the examination of detection error trade-off (DET) curves (Martin et al., 1997). These curves represent the operating characteristics of the system through a plot of the miss probabilities of the system (the probability of rejecting the utterance as the correct language) against the false alarm probabilities (the probability of accepting the utterance as the wrong language). The primary measure used in this study to compare any two language ID systems is the equal error rate (EER). The EER is the performance of the system when the miss probability equals the false alarm probability. In order to obtain the impostor and target scores required to derive these curves, normalisation had to be performed on the identification scores. Individual language scores for each utterance were normalised by calculating a likelihood ratio score. The ratio is calculated by dividing the language score by the sum of the other eleven language scores for that utterance. Where appropriate, language identification rates (percentage of correct identifications) are also given.

4. Syllable length framework

In this study, phone-triplets are used to provide a segmentation framework for subsequent model development for the language identification task. This segmentation is achieved by recognising broad phonetic classes (BPC) using a multilingual broad phone recogniser, and then concatenating these phones to form phone-triplets or pseudo-syllabic events.

The overall system operation is depicted in Fig. 1. The front end phone recognition system produces a sequence of broad phonetic events. The number of broad phonetic classes was limited to four in order to restrict the total number of possible syllables. Further details of the recognition system and class definitions are outlined in Section 5. The broad phonetic transcription is subsequently converted into a transcription containing broad class phone-triplets, which acts as a crude representation for the syllabic event. These events are referred to as *pseudo-syllabic* because they do not necessarily reflect the process expected from true syllabic segmentation and the subsequent

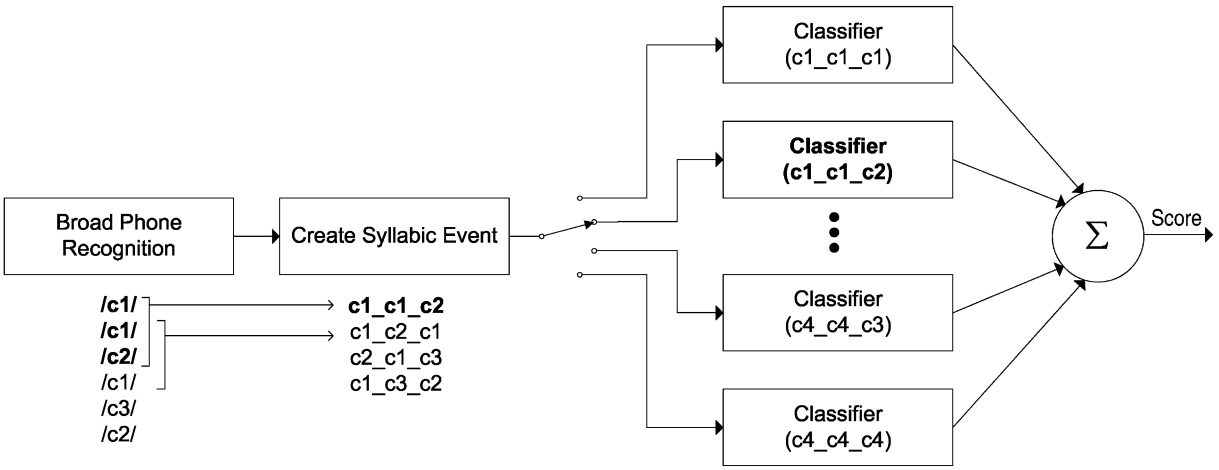

Fig. 1. Syllable-length framework.

boundary information this would provide. Throughout the remainder of this study, the set of broad syllabic events is denoted by ψ.

Given that each pseudo-syllable contains three phones, and the number of possible broad phonetic classes is four, the resulting number of syllables in the set ψ is 64. This small set size ensures that sufficient training data are available for each syllable across all languages. It should be noted that in time stamping each instance of ψ, overlapping windows were used. That is, a window of three phones was advanced one phone at a time. It is expected that trajectory information which over-arches phonetic events is useful for discrimination across languages. Accordingly, by progressing the time-stamping for ψ one phone at a time, trajectory information can be exploited that occurs across traditional syllable boundaries. In contrast to true syllabic segmentation, this method also provides more instances of training data for each sequence of three phones and importantly more assessment information during testing, which is particularly important when testing shorter utterances. However, it is acknowledged that the use of overlapping windows may result in loss of important information contained within true syllabic boundaries.

The boundary information obtained via the segmentation process is then used to extract features and train individual classifiers for each 3-phone event. In this way, a classifier is available for each syllable and its success can be examined in isolation or in conjunction with other syllabic classifiers. Fig. 1 illustrates that a score for a particular language can be obtained by simply summing the outputs obtained by the appropriate individual classifier, over the duration of the utterance. This score is then compared to scores obtained for each of the languages and the score with the highest likelihood selected. Section 4.1 provides a mathematical formulation for this identification process.

4.1. System formulation

Let each utterance be represented by a series of feature observations defined by $O = (o_1, o_2, \ldots, o_p)$, which is subsequently mapped to a sequence of broad class phonetic phone-triplets $B = (b_1, b_2, \ldots, b_m)$. This is achieved using a phone recognition system. Each of the entries in the stream B, is one of $N = 64$ possible entries defined by the lexicon $D = (\psi_1, \psi_2, \ldots, \psi_N)$.

For each language L, a set of models is produced for the entries in D, and denoted using I_L^{ψ}, where I can be any of the proposed acoustic, phonotactic or prosodic information sources. These models can then be used to evaluate the language identity for each test utterance according to:

$$L_{\mathrm{ID}}^{I} = \arg \max_{L=1\ldots W} \sum_{j=1}^{m} \sum_{k=1}^{T_j} \log p(o_k | b_j, I_{L_W}^{\psi_j}) \tag{1}$$

where W denotes the number of languages and T_j refers to the number of observations associated with the recognised triplet b_j.

5. Front-end recogniser

In an attempt to obtain the most accurate segmentation, two separate multilingual phone recognition systems were trialled. Both systems used context independent HMMs with a 3 state

left-to-right model topology. Speech was parameterized using 12th order PLP analysis plus normalised energy, 1st and 2nd order derivatives and a frame size/shift of 25/10 ms, respectively. Utterance based cepstral mean subtraction was employed.

The first recognition system incorporates a model set containing 20 multilingual phonetic groups, with each state emission density modelled using 32 mixture components. These groups were selected based on providing broad articulatory coverage for the range of sounds likely to occur across multiple languages, in tandem with experience gained by using decision trees to cluster multilingual phone sets (Wong et al., 2003). Given the decision to restrict the number of broad syllabic events to 64, this meant that these multilingual phones had to be subsequently mapped back to one of four possible broad classes. Table 2 provides details for the 20 models and the corresponding mapped substitutes. The second recognition system was trained to directly model the four acoustic classes, outlined in the first column of Table 2, with state emission densities modelled using 256 mixtures.

The final four broad classes were selected based on the work conducted by Kajarekar (2002). In this work, Kajarekar examined the speaker and channel variability of phonetic classes. He highlighted that the *F*-ratio is a common formulation used to maximise class separability and used this as a discriminatory measure to establish the following four broad phonetic classes:

Table 2
Multilingual phone set and broad phonetic groupings for the front end recogniser

Broad phonetic class (C)	Articulatory description for models (M)	Worldbet examples
Vowels and dipthongs (c1)	m_1 – i type Vowels	*I*, *Ix*, *If*, *y*
	m_2 – u type Vowels	*u*, *U*, *u:*
	m_3 – o type Vowels	5, >, *o*
	m_4 – e type Vowels	*E*, 3*r*, 7, 8
	m_5 – a type Vowels	*a*, @, *A*
	m_6 – Dipthongs	*ei*, *aI*, *iU*, *aU*
Nasals, liquids, and glides (c2)	m_7 – Nasals	*m*, *n*, *n* = *nr*
	m_8 – Approxs. and taps	*r*, *rr*, *r*{*H*, 9*r*
	m_9 – Lateral alveolar approx	*l*, *L*, *l :*
	m_{10} – Palatal approx	*j*
	m_{11} – Voiced lab. velar approx	*w*
Fricatives (c3)	m_{12} – Unvoice frics. and affrics	*tS*, *f*, *s*, *S*
	m_{13} – Voiced frics. and affrics	*D*, *dZ*, *G*, *z*
Silence and stops (c4)	m_{14} – Silence, pause and noise	+, #, *.ls*, *.ln*
	m_{15} – Voice bilabial plosives	*b*, *bH*
	m_{16} – Unvoiced bilabial plosives	*p*, *pH*
	m_{17} – Voice alv. and retro plos.	*d*, *d*{, *dH*
	m_{18} – U/V alv. and retro. plos.	*t*, *tH*, *t*{*H*
	m_{19} – Voiced velar plosives	*g*, *gH*
	m_{20} – Unvoiced velar plosives	*k*, *kH*

Table 3
ID rate and equal error rates using two broad phonetic class (BPC) front-end recognisers

Front end recogniser	30 s Test (ID Rate)	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
20 Phone → 4 BPC	66.2%	15.6%	18.6%	23.8%
4 BPC	65.9%	16.1%	18.6%	24.2%

- vowels and dipthongs (c_1),
- nasals, liquids and glides (c_2),
- fricatives (c_3),
- silence and stops (c_4),

By modelling these individual groupings using GMMs, Kajarekar obtained superior performance in speaker recognition experiments, when compared to the traditional approach that models all active speech using a single GMM.

The two front end systems produced hypothesised transcriptions that exhibited a number of differences, including different segment boundaries, and variations in the number and types of substitutions, insertions and deletions. Initial training attempts achieved similar recognition performance for both systems (after mapping back to four classes); however after introducing skip states to the silence/pause model of the 20-phone system, a recognition accuracy of 69.0% was achieved. In comparison the four model system only achieved 61.0% recognition accuracy. This level of recognition accuracy means that the subsequent phone-triplet models will be trained using sub-optimal transcriptions. However, phonotactic systems currently used in LID systems are based around phone recognition systems which typically have 40–60% error rates and still perform quite well.

Whilst improved segmentation accuracy is desirable, the more important consideration is whether it translates to improved LID rates. Accordingly, an additional experiment was conducted that used the transcriptions produced by both recognition systems and assessed their accuracy in a LID task on the CallFriend database. The syllabic transcriptions produced by both front ends were subsequently used to train a classifier for each of the 64 events, using 64 Gaussian mixture components to model the density. The language identification performances for the two systems are shown in Table 3.

Whilst the results in Table 3 indicate only marginally better performance across the three testing times, the 20 class system had the added benefit that the transcriptions could subsequently be used to map back to any number of final broad classes (aside from the currently used four). Given this, the 20 class system was used for recognition purposes in all other reported experiments.

6. Acoustic modelling using the syllable-length framework

This section outlines a comparison of techniques used for modelling acoustic features within the syllable-length framework. The first comparison, outlined in Section 6.1 and 6.2, compares a baseline GMM system with a system producing GMM models for each of the phone-triplet segments. Results of the comparison are provided in 6.2. A further comparison is then provided for a HMM based system against the GMM systems in Section 6.3. Reported results are EERs for the 3, 10

and 30 s test utterances from the CallFriend database, according to the NIST 1996 protocol as outlined in Section 3.

6.1. Baseline acoustic modelling using GMMs

A GMM/UBM (Reynolds, 1997) language identification system, based on the work outlined in Wong and Sridharan (2003), was implemented as a baseline for performance comparisons. The GMM/UBM system used 16 dimension feature vectors, which included 5th order perceptual linear predictive (PLP) coefficients (Hermansky, 1990), delta, and acceleration values, as well as delta energy coefficients. Mean and variance normalisation was applied to all feature vectors to reduce the noise and channel effects. Vocal tract length normalisation, as described in Wong and Sridharan (2003), was also incorporated.

The UBM used was a 512 component GMM, trained using data from all languages. Models for each language were then derived using Bayesian adaptation. The results achieved by this system are discussed and compared in Section 6.2. This system is denoted using GMM_{512Mix}.

6.2. Acoustic modelling using the syllable-length framework and GMMs

One of the aims of this study was to re-examine whether benefit can be obtained by using the syllable-length segmental unit to subdivide the feature space prior to acoustic modelling. In order to evaluate this idea, a GMM/UBM modelling technique was used to model the acoustic feature space for each of the possible syllabic events. The same PLP features outlined in Section 6.1 were used.

In accordance with Eq. 1, each syllabic event is designated ψ_j, where j can be one of 64 possible models. The UBM models for each event ψ_j were 64 component GMMs (as opposed to 512 used for the baseline), created using data from all languages. Models for each language L, were then derived using Bayesian adaptation and designated as γ_L^{ψ}. By substituting γ for the general form I in Eq. 1, the language identity can be determined.

Table 4 provides a comparison between the GMM baseline system GMM_{512Mix}, and the system using syllabic segmentation $GMM(\psi)$. Confidence intervals are also provided at the 95% level of statistical significance for these systems, and all other systems quoted throughout the remainder of this paper are based on the same level of statistical significance.

Examination of Table 4 shows that the systems achieve almost identical levels of performance, indicating little benefit is obtained using Gaussian mixture modelling of phone-triplet segmental units. Fig. 2 illustrates the differences between the two systems when evaluated on the 30 s test.

Table 4
Equal error rates for the baseline and syllable-based GMM acoustic systems

Modelling description	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
$GMM_{512\text{Mix}}$	$15.5\%^{+1.4}_{-1.1}$	$18.6\%^{+1.4}_{-1.1}$	$23.8\%^{+1.5}_{-1.3}$
$GMM(\psi)$	$15.6\%^{+1.5}_{-1.3}$	$18.6\%^{+1.2}_{-1.3}$	$23.8\%^{+1.4}_{-1.3}$
$GMM(\psi_{\text{Top32}})$	15.7%	18.8%	24.3%
$GMM(\psi_{\text{Top16}})$	16.3%	19.6%	25.3%
$GMM(\psi_{\text{Top8}})$	17.0%	21.0%	26.2%

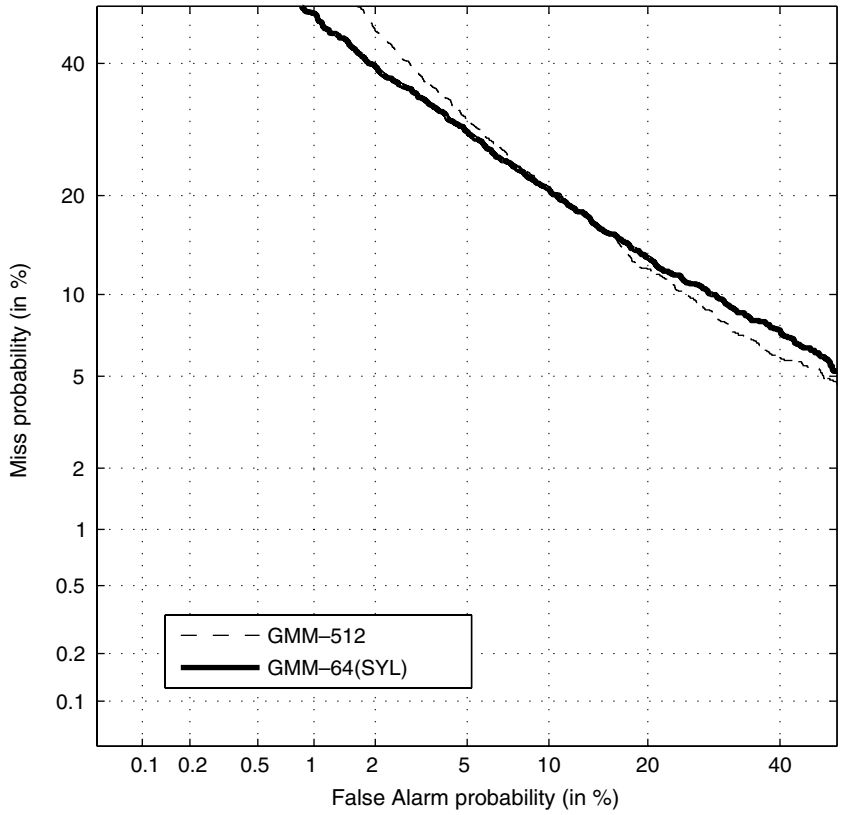

Fig. 2. DET plot comparing baseline acoustic GMM system and syllable-length framework GMM.

This highlights the similarities in performance, but the improvement seems quite pronounced in the low FA region.

As mentioned in Section 2, the use of SDC produced state-of-the-art levels of performance in modelling short term acoustic features, but little linguistic insight can be obtained using this method. Interpreting the results of the segmentation process more optimistically, using the segmental framework did not result in a degradation in overall acoustic modelling performance, but does provide scope for analysing the discriminatory capability for each of the segmental units. Accordingly, individual performance measurements were gathered for each syllabic event, tested in isolation. These EER are then cross-compared with individual frequency of occurrence rates and a graphical version of this information is provided for the 32 best performing syllables in Fig. 3.

Fig. 3 indicates that a definite correlation between frequency of occurrence and EER exists. Whilst it is possible for syllables which occur infrequently to have low EER's, the usefulness of these syllables are limited because they fail to occur with sufficient frequency to contribute significantly over an entire utterance. However, the general correlation indicates that EER can be used as an effective representative for the discriminative capability of each syllable. Accordingly, individual syllable EER statistics were recalculated using the NIST1996 development data, and EER was then used as a selection criterion for choosing the best performing subsets of 8, 16 and 32

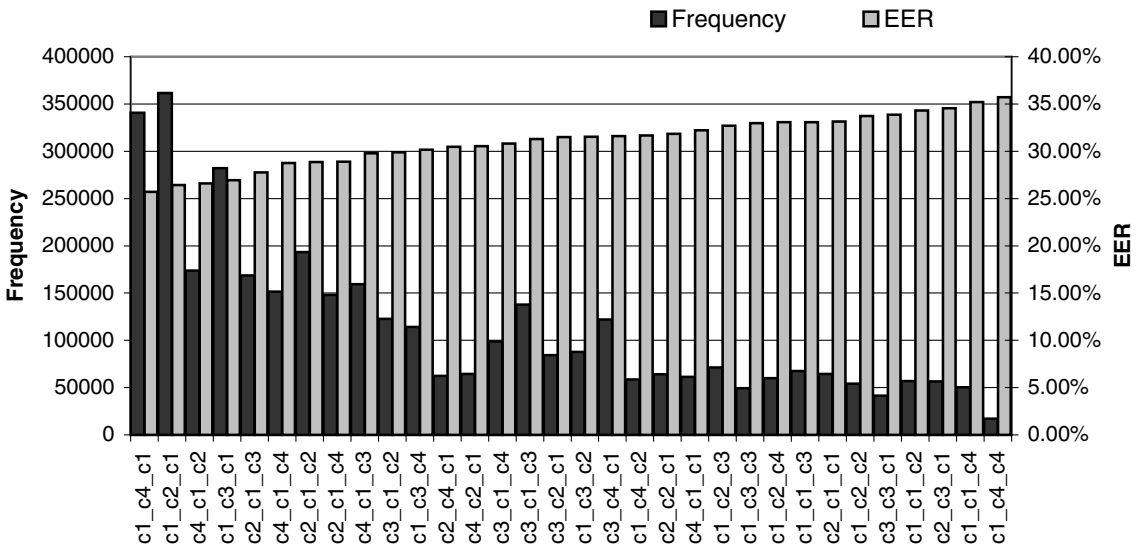

Fig. 3. Frequency of occurrence and EER performance of syllabic-events modelled using GMMs of acoustic features.

syllabic events. These subsets were then re-evaluated on the test set and the results obtained are contained in Table 4. It can be seen that even when only 8 of the possible 64 segmental units are used, the level of degradation is minor. An examination of the top eight indicates that the triplets contain at least one vowel and generally a nasal. This is a good indication that the system could probably benefit from increased resolution of these events. In contrast, it also indicates that little benefit can be obtained by increasing the phonetic resolution for fricatives, stops and silence.

6.3. Acoustic modelling using the syllable-length framework and HMMs

In contrast to Section 6.2, this section outlines the use of HMMs for modelling the phone-triplets. The feature set modelled replicates those used previously for the GMM modelling. Each syllabic event ψ is modelled by producing a set of HMM's λ for each of the 64 possible syllabic events.

A universal set of syllabic HMM's was first built using the segmentation information produced by the front end recogniser for all training languages. Maximum a priori (MAP) adaptation was then used to produce language specific models denoted by λ_L^{ψ}. A 9-state left to right state topology was chosen after experimenting with both 5 and 9 states. Additionally, a number of experiments were conducted in an attempt to establish the most suitable adaptation factors, although these were not exhaustive. 32 mixture components were used to model each state distribution.

Table 5 summarises the performance of the HMM based system, again with significance figures provided for the basic system. For ease of comparison, the corresponding syllable based GMM system, which performed almost identically to the baseline GMM system, is also included. Fig. 4 also provides a comparison between the GMM and HMM based syllable systems when evaluated on the 30 s test. Initial inspection of Fig. 4 indicates that HMM modelling provides a general improvement, however further examination of Table 5 reveals that the differences are not significant at the 95% level of confidence.

Table 5
Equal error rates for the syllable-based HMM acoustic system

Modelling description	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
$GMM(\psi)$	$15.6\%^{+1.5}_{-1.3}$	$18.6\%^{+1.2}_{-1.3}$	$23.8\%^{+1.4}_{-1.3}$
$HMM(\psi)$	$13.8\%^{+1.4}_{-1.1}$	$16.7\%^{+1.4}_{-1.2}$	$22.5\%^{+1.7}_{-1.4}$
$HMM(\psi_{6\text{Class}})$	$12.3\%^{+1.2}_{-1.0}$	$15.4\%^{+1.3}_{-1.2}$	$21.5\%^{+1.6}_{-1.4}$
$HMM(\psi_{\text{Top32}})$	13.9%	16.8%	22.5%
$HMM(\psi_{\text{Top16}})$	14.5%	17.6%	23.5%
$HMM(\psi_{\text{Top8}})$	14.8%	18.3%	24.8%
HMM_{phone}	26.4%	29.4%	33.9%

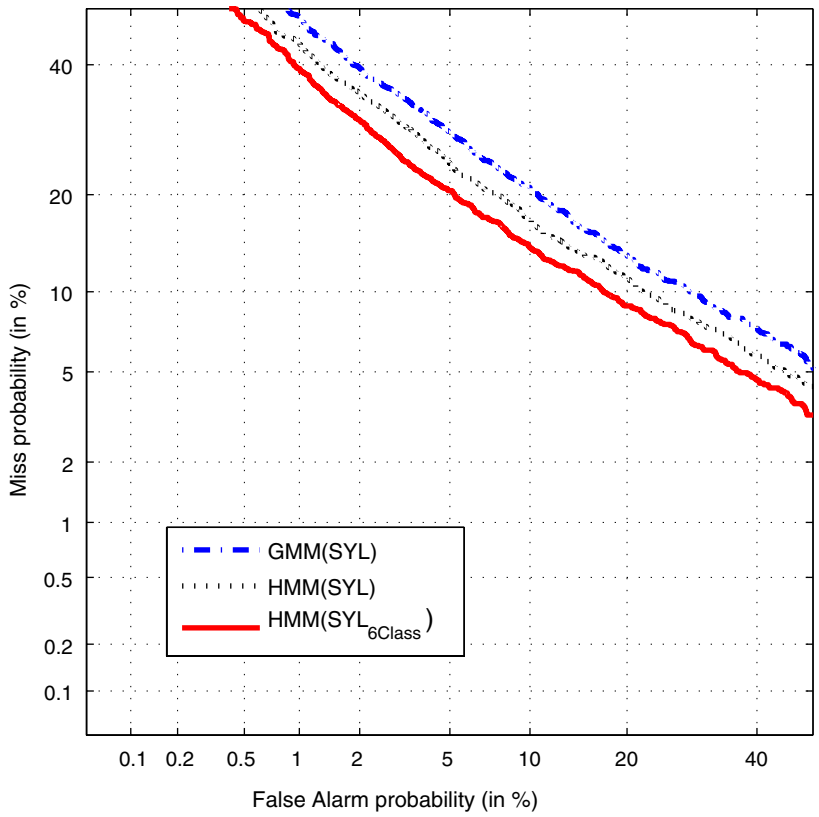

Fig. 4. DET plot comparing baseline syllable-based acoustic GMM and HMM systems for the 30 s test.

Once again the development set was used to determine the top performing syllables, with the top 32, 16 and 8 then evaluated on the test set; these are included in Table 5. It can be seen that most of the discriminative contribution comes from the top eight syllables. As in Section 6.2, a plot of EER vs. frequency of occurrence was produced; however, the results obtained using the HMM virtually mirror those for the GMM system, and so this figure was not reproduced.

In order to assess whether the use of an extended segmental unit was providing any benefit over the more commonly used phone-length segmental unit, an additional experiment was conducted.

In this experiment, the HMM system using the phone-triplets, $HMM(\psi)$, was compared to a system which used the four broad phonetic classes as the segmental unit, HMM_{phone}. This result is included in Table 5 and it can be clearly seen that the extended unit is superior by a considerable margin.

It has been stated on several occasions that the framework can be used to extract meaningful insight regarding which syllables contribute the most to overall language discrimination. In an attempt to exploit this insight, and examine its *practical* usefulness, the set of four broad class phones was expanded to 6, using the information gleaned from Fig. 3. Given vowels existed in all of the syllables with more discriminatory power, the original vowel class, (c_1) was expanded to three new classes. The new classes were selected using a binary decision tree process, similar to that used for clustering context-dependant phones as outlined in Young et al. (2002). Using a starting cluster containing all vowels and dipthongs in the 20 class set, the tree was grown and subsequently examined to select the classes. The groupings which resulted, based on the entries in the second column of Table 2, were (m_4, m_6), (m_3, m_5) and (m_1, m_2). Table 5 also includes results for the system using six classes to build the triplets, $HMM(\psi_{6\text{Class}})$, with each event modelled using a HMM topology. It can be seen that this results in improvements ranging from 1.5% EER for the 30 s test, to 1.0% for the 3 s test, over the original HMM based system. Of further note is that the 6 Class system provides statistically significant improvements to the original GMM based system, GMM_{512}, for the 30 and 10 s tests.

7. Phonotactic modelling of syllabic events

As highlighted in Muthusamy (1993), languages have different characteristic sound patterns, which can be described subjectively using terms such as *singsong*, *rhythmic*, *nasal* or even *guttural*. For instance, a language such as Italian can be described as melodic because of its high vowel-to-consonant ratio within each syllable, whereas Mandarin has a high occurrence of nasal sounds. Alternatively, Japanese has a quite strict consonant–vowel syllabic structure, whereas German and English (Berkling, 1996) allow concatenated sequences of several consonants. This rhythm may be more effectively captured by examining broad phone clusters, an effect which may not be captured using the fine phonetic resolution used by PPRLM systems. It has been suggested that the syllable is the natural unit of rhythm, and accordingly capturing frequency statistics for each broad syllabic event ψ may capture this rhythmic information. Accordingly, these unigram statistics were gathered and the accumulated likelihood scores of the test utterances used to determine the language. Effectively, this equates to scoring the phone sequences produced by the broad phone recogniser using a *Bag-of-N-Grams* classifier, first introduced in Doddington (2001), with a n-gram length of $n = 3$.

Two different syllabic unigram systems were trialled. The first modelled the unigram statistics of the syllabic events produced by concatenating into triplets the four recognised (after mapping), broad phone classes. The second system used the original 20 multilingual phonetic descriptions, again reformatted to form phone-triplets. A comparison of these two systems illustrates the impact that increasing (or decreasing) levels of phonetic resolution have on identification performance. In order to combat model sparsity issues, a smoothing of the unigram models produced by the 20 phone system was performed by using the MAP adaptation process outlined

Table 6
Equal error rates for the syllable-based unigram and PPRLM system

Modelling description	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
Unigram($\psi_{4\text{phn}}$)	26.1%	31.0%	37.4%
Unigram($\psi_{20\text{phn}}$)	10.5%	17.2%	27.2%
PPRLM	6.4%	10.7%	19.0%

in Baker et al. (2004). This adaptation process was also trialled for the four broad phone systems, but was found to provide no benefit.

Table 6 shows the resulting performance of the two unigram modelling systems. It can be seen from the results that the increased phonetic detail provided by the 20 phone system gives far superior performance to that of the four phone system. However, the 26.09% EER obtained by the four phone system for the 30 s test is still commendable, considering the system uses only 64 probability statistics for each language (compared to 8000 used in the 20 phone system, and tens of thousands used in a traditional PPRLM based system). These results highlight the usefulness and importance of modelling phonetic sequences and patterns for distinguishing between languages, and reinforce the findings of previous studies in this area (Zissman and Singer, 1994).

It is also interesting to compare these modelling techniques against, and in combination with a traditional PPRLM LID system (Zissman and Singer, 1994). PPRLM has been shown to be one of the most effective strategies for LID, particularly for longer length test utterances. A PPRLM system constructed at QUT was used for comparison with the unigram systems. This system uses six open loop phone recognisers, each trained to recognise phonetic events for a different language. The languages used are English, German, Hindi, Japanese, Mandarin, and Spanish, with the open loop phone recognisers trained using data from the OGI-MLTS database. In contrast to the unigram (or bag of 3grams) models used for the syllabic systems, the trigram-based PPRLM system incorporates conditional probabilities, rather than joint probabilities used in the *bag-of-ngram* system. Further details on the construction and previous performance of this particular PPRLM system can be found in Wong and Sridharan (2003). A more detailed explanation of the general PPRLM framework is described in Zissman and Singer (1994). Table 6 shows the performance of the PPRLM system compared to the unigram syllabic systems described above. As expected, the finer phonetic resolution provides definite benefits in terms of LID performance.

Experiments were also performed in order to determine whether the phonetic information provided by the syllable-length framework could provide complementary classifications to those provided by the PPRLM system. A multilayer perceptron (MLP) neural network, implemented

Table 7
Equal error rates for fused combinations of the syllable-based unigram systems and PPRLM system

Modelling description	30 s Test (EER)
$U(\psi_{4\text{phn}})$-Triplet unigram-coarse	23.1%
$U(\psi_{20\text{phn}})$-Triplet unigram-fine	9.3%
PPRLM-QUT 2003 NIST System	4.9%
$\text{U}(\psi_{4\text{phn}}) + \text{U}(\psi_{20\text{phn}})$	9.5%
$\text{PPRLM} + \text{U}(\psi_{4\text{phn}})$	4.4%
$\text{PPRLM} + \text{U}(\psi_{20\text{phn}})$	3.4%
$\text{PPRLM} + \text{U}(\psi_{4\text{phn}}) + \text{U}(\psi_{20\text{phn}})$	3.9%

using the *LNKnet* package was used (Massachusetts Institute of Technology Lincoln Laboratory, 2004). Further details of the methodology used for fusion are outlined in Section 9. Table 7 contains the results for these experiments when the 30 s test segments were used. Results are provided for each system used in isolation (but fed through the MLP), as well as various combinations of the four phone, 20 phone and PPRLM systems.

It should be noted that differing EER performances for the individual systems can be found when comparing the results in Tables 6 and 7. These differences exist because of the non-linear combination of scores that the neural network performs. The MLP is given the 12 language scores as a feature vector for each test utterance. The MLP is then able to exploit patterns found (if present) in the distribution and pairings (or larger combinations) of these scores. That is, by using the MLP, the individual language scores are no longer treated independently. This non-linear combination, in some cases, leads to improvements in performance of individual systems. Accordingly, it was necessary to obtain performance measures for each individual modelling paradigm using the neural network, so that an equitable comparison can be conducted between the results of individual systems and those obtained by combining multiple systems.

Fig. 5 shows DET curves for the most interesting of these fusion combinations. From this plot, it can be seen that a considerable improvement in performance is achieved by fusing the PPRLM system with the 20 phone unigram system. The addition of this unigram modelling of the 20 phone

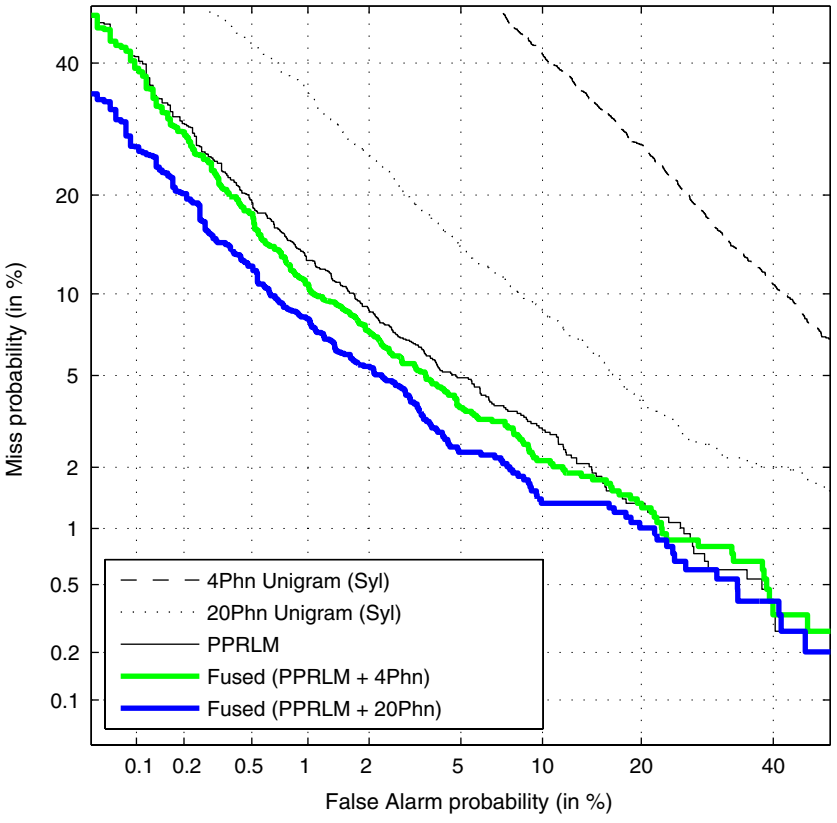

Fig. 5. DET plot showing performance of a fused PPRLM and syllable-based unigram system for 30 s test.

syllables gives a 30% relative improvement over the PPRLM system, clearly indicating the utility of the extended segment statistic.

However, only a slight performance gain is achieved by adding the four phone system to the PPRLM system. More importantly, it can be seen in Table 7 that no performance gain is obtained by combining both the 20 and 4 phone syllabic unigram systems with the PPRLM system. This suggests that no information exists in the coarse modelling approach, which cannot be extracted from the system with greater phonetic resolution, supporting Muthusamy's suggestion (Muthusamy, 1993).

8. Pitch and energy modelling

As mentioned in Section 2, several studies have been conducted which make use of prosodic information. However, many of these studies have examined the use of prosodic information in isolation, or when combined with other information sources were examined on smaller language subsets. Accordingly, we outline a system produced under the common syllabic framework, and compare its performance with a system implemented in accordance with the basic principles outlined in Adami and Hermansky (2003), using the NIST 1996 protocol. The system based on Adami's techniques is outlined in Section 8.1, whilst the proposed technique and appropriate comparisons and results are provided in Section 8.2.

8.1. Baseline pitch and energy modelling using N-grams

In Adami and Hermansky (2003), a small set of descriptive tokens was used to represent stylized pitch and energy segments. Simple *n*-gram models (*bag-of-n-gram* models) of the token sequences were used to determine the language. The relatively good performance obtained given a small, compact and rudimentary feature set suggested that the temporal trajectories of pitch and energy provide a reasonably robust technique for language identification.

A pitch and energy modelling system, based upon the technique proposed by Adami and Hermansky (2003), was implemented as a baseline for performance comparisons. In the baseline system, prosodic features are estimated by representing the pitch and energy contours for an utterance as a series of piecewise linear segments and subsequently identifying each of these segments as a single prosodic event. Each event is characterised in terms of the slope of the pitch contour (positive or negative), the slope of the associated energy contour, and by its duration (short or long). Unvoiced events are quantised simply in terms of their duration, resulting in a total of 10 descriptive tokens. The boundaries and labels of the four broad phone class transcriptions were also used to further segment the data, resulting in a total of 40 possible classes. This differs from Adami's implementation (Adami and Hermansky, 2003), which used six broad phonetic categories, but ensures that the comparison between systems is conducted on a more equitable basis.

To obtain the piecewise linear segments, the following steps were followed. Firstly, the $f0$ and short-time energy values were computed every 10 ms. The $f0$ values were obtained using the *getf0* utility (Talkin, 1995), which implements a pitch tracking algorithm based on the use of the cross correlation function and dynamic programming. The time derivatives for both the $f0$ and energy trajectories were then calculated by fitting a straight line to 10 frames (100 ms). The speech was

then segmented using the zero-crossings of this rate of change values and broad phone boundaries. Segments were then labelled according to the slope of the two trajectories, the current phonetic category, and the length of the segment (segments shorter than 8 frames were labelled "*Short*" and all other segments were labelled "*Long*").

N-gram models, similar to the one described in Doddington (2001), were used to produce language dependent models. An *n*-gram length of $n = 3$ was found to be optimal in both our experiments and those performed in Adami and Hermansky (2003).

8.2. Pitch and energy modelling using the syllable-length framework and HMMs

Given that Adami and Hermansky (2003) achieved respectable results using a limited set of description based tokens to describe pitch and energy trajectories, it was decided to test whether these trajectories could be better modelled making use of the syllable-length framework and using a HMM topology. Given the expected variation in duration, the HMM offers a form of duration normalisation by incorporation of this variation in transition probabilities. Additionally the HMM provides a soft quantisation of trajectory events, as opposed to a tokenised description such as "+*ve*" or "−*ve*".

A HMM was used to model both pitch and energy trajectory information for each broad syllabic event ψ_j where $j = \{1 \ldots N\}$ and $N = 64$, with the models for each event from each language denoted $\lambda^{\psi}_{pitch_L}$. By substituting λ_{pitch} for the general form I in Eq. 1, the language identity can be determined.

The feature vectors used consisted of delta pitch and energy values, estimated every 10 ms. These pitch values were calculated using a normalised cross-correlation technique with subsequent delta values extracted by fitting a straight line over 9 frames. The log of the energy was calculated before delta values were calculated. Some initial experiments were performed using various window sizes for the delta calculations, with a length of 9 frames found to be optimal.

Initial inspection of trajectories indicated that 5 states would be sufficient to capture coarse absolute changes in direction of the pitch and energy trajectories, with the mixture components within each state distribution density modelling the variation between these absolute changes.

In a previous study (Martin et al., 2004), language dependent models were developed with 4 mixture components assigned to each state density distribution. Subsequent improvements to this technique have been achieved by adapting language dependent models from a background model trained using data from all languages. This adaptation process allowed for an increase in the number of mixture components to 32.

Results were obtained for both the baseline pitch and energy system (which used tri-gram modelling of the descriptive tokens), and the new syllable-length framework HMM based system. Table 8 provides a comparison of the results for these two systems in terms of EER for the 3, 10 and 30 s tests. Examining these results, it can be seen that no significant gain is achieved over the *n*-gram baseline. Comparable performance was achieved for the 30 and 10 s tests, and only a small improvement over the baseline was achieved for the 3 s trials. This slight improvement for the 3 s trial may suggest that the HMM based approach to modelling the trajectories is more robust for shorter durations; however significance testing on these results shows that such a claim cannot be made without further testing and more definitive results.

Table 8
Equal error rates for the baseline and HMM pitch and energy systems

Modelling description	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
Pitch baseline	$26.2\%^{+1.2}_{-1.2}$	$29.4\%^{+1.6}_{-1.3}$	$35.5\%^{+1.7}_{-1.6}$
Pitch HMM(ψ)	$27.8\%^{+1.6}_{-1.6}$	$29.6\%^{+1.6}_{-1.5}$	$33.5\%^{+1.8}_{-1.8}$
Pitch HMM(ψ_{6Class})	$27.4\%^{+1.6}_{-1.6}$	$29.4\%^{+1.5}_{-1.5}$	$33.9\%^{+1.9}_{-1.8}$
Pitch HMM(ψ_{top32})	27.7%	29.8%	34.2%
Pitch HMM(ψ_{top16})	27.5%	29.4%	34.1%
Pitch HMM(ψ_{top8})	27.7%	31.3%	35.6%

The syllable-length framework also allowed for further analysis of the results obtained from this system. In a similar manner to the analysis performed in Section 6, a crude indication of the discriminative capability of the individual syllable HMMs can be obtained, by analysing the performance when tested in isolation. The EER was computed for each syllable in isolation, with resulting values ranging from 35% to 47% for the 30 s trial. Fig. 6 shows the frequency count, along with the EER for the top 32 syllabic events (in terms of EER) when used as isolated classifiers. In contrast to the correlation effect outlined for the acoustic system in Section 6.2, there is a less distinct correlatory relationship between the frequency of the event and EER performance. Inspection of the first 12 syllables reveals some correlation; however the trend is less distinct when the set is considered in its entirety.

Subsequent investigation of the internal structure of the higher performing syllables revealed similar trends to those revealed in Section 6, with 5 of the top 8 syllables coinciding with the top 8 for the acoustic based system. In fact, 9 of the top 12 syllables (in terms of EER) contained at least one nasal/glide and one vowel/dipthong. This reinforces work in Kajarekar (2002), and

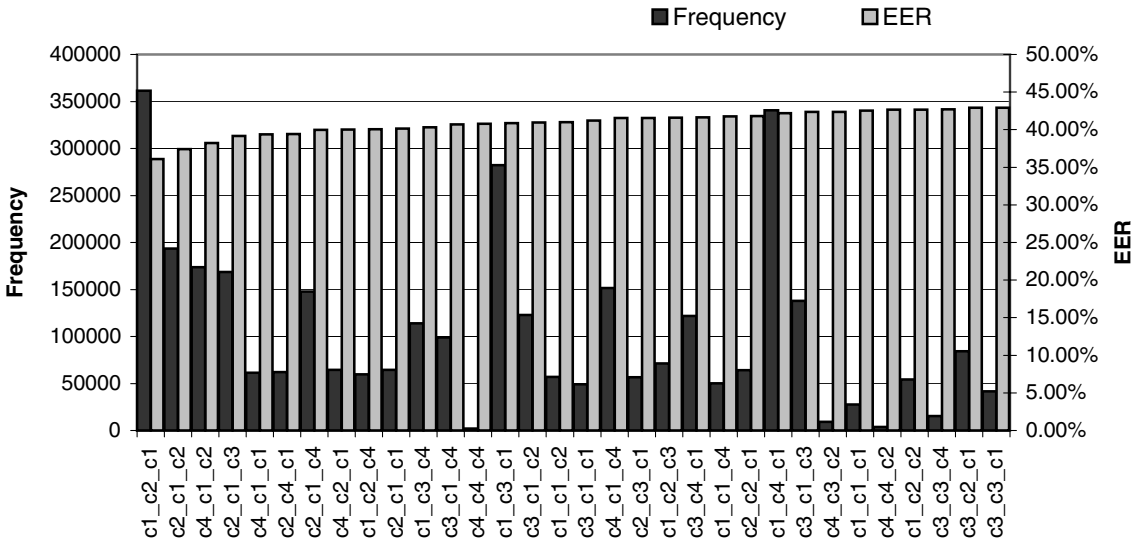

Fig. 6. Frequency of occurrence and EER performance of syllabic-events modelled using HMMs of pitch and energy deltas.

highlights that these phonetic groups form the more discriminatory segmentation units. As found with the acoustic systems, those syllabic events that contained only one or no instances of these two broad phonetic classes performed poorly. This may also be an indication of the greater degree of voicing information contained by the phones grouped within the vowel/dipthong and nasal/glide classes.

In a similar manner to Section 6, the best performing subsets of 8, 16 and 32 were calculated using the development data and then evaluated on the test data. The results for these tests are contained in Table 8. Interestingly, reducing the syllable set for the pitch system actually gave some slight performance improvements in some cases. The best performing syllabic set for the 30 and 10 s trials was actually the top 16 set. This improvement is most likely due to the pruning of the more erroneous, unvoiced syllables with high frequency of occurrence.

As a side experiment, it was decided to fuse the baseline pitch classifications with those from the HMM based system in order to determine if there is any complementary information produced by these two techniques. The fusion was performed using a MLP neural network implemented in LNKnet (Massachusetts Institute of Technology Lincoln Laboratory, 2004). Further details of the procedure used in fusion experiments can be found in Section 9. Fig. 7 shows DET curves for the 30 s test condition comparing the baseline system, the HMM system using all 64 syllables, and the fused combination of these two systems. The plot clearly shows that a significant gain can

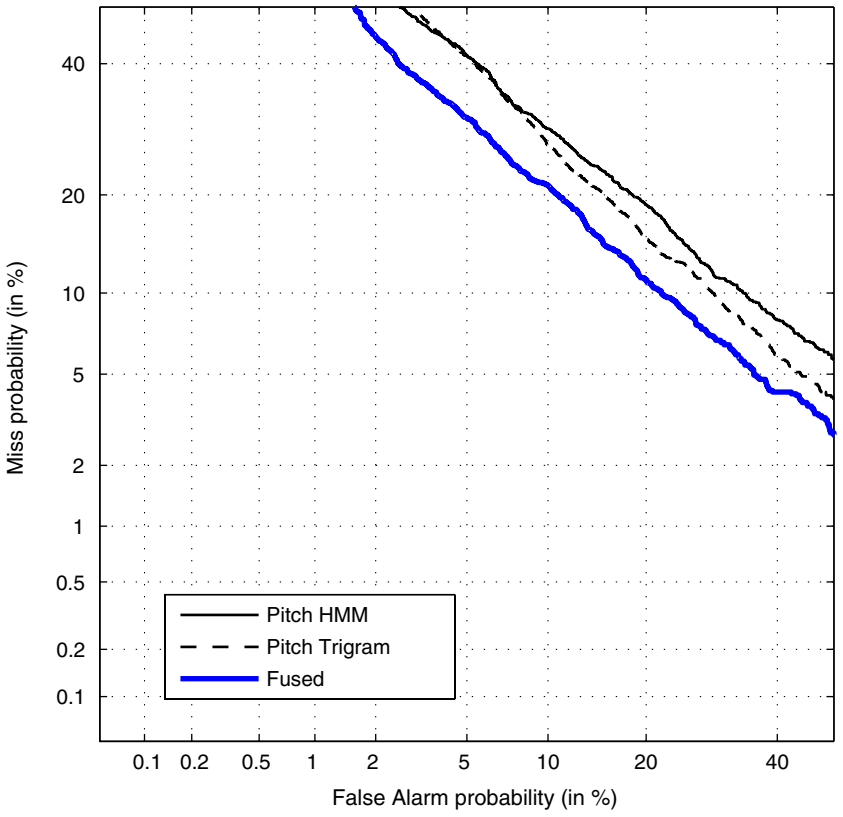

Fig. 7. DET plot for pitch and energy systems with fusion results for the 30 s test condition.

be achieved through fusion of the two systems. The fused system achieved an EER of 14.5%, which was an absolute improvement of 4.8% over the original HMM based system, and an absolute improvement of 3% over the system in the style of Adami's.

Finally, in a similar manner to Section 6.3, the expanded six class phoneset was used to produce a new set of pitch based syllabic models and is designated Pitch HMM($\psi_{6\text{Class}}$) in Table 8. Disappointingly, the increased resolution did not provide any significant improvement and actually degraded slightly for 3 s test.

9. Overall results and fusion

As the previous sections have mentioned, language identity is encapsulated in a complex manner. No individual feature set provides a panacea for the task of language identification. As such, fusion of multiple systems represents an approach for improving performance through the extraction of complementary information.

To determine whether the features and associated modelling techniques outlined in this study are in fact complementary, a simple set of fusion experiments were conducted. A MLP neural network was constructed using the LNKnet pattern classification software (Massachusetts Institute of Technology Lincoln Laboratory, 2004). The neural network had 50 nodes in the hidden layer, and experiments were performed using a 5-fold cross validation procedure on the NIST1996 evaluation test set.

As outlined in Section 7, differences exist between the scores for individual system obtained before and after processing via the neural net. Table 9 contains the identification rate and EER's for the individual systems after this process, enabling more equitable comparisons and also acts to consolidate results achieved for the individual systems.

Table 9
ID rate and equal error rates for individual systems after non-linear fusion of scores using a MLP

System designator	Description	30 s Test (ID Rate)	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
$\lambda(\psi)$	Acoustic HMM	$80.4\%^{+1.9}_{-1.9}$	$8.0\%^{+1.1}_{-1.0}$	$12.6\%^{+1.4}_{-1.1}$	$20.6\%^{+1.4}_{-1.5}$
$\lambda_{\text{pitch}}(\psi)$	Prosodic HMM	$55.0\%^{+2.0}_{-2.1}$	$19.3\%^{+1.4}_{-1.5}$	$24.2\%^{+1.6}_{-1.4}$	$30.3\%^{+1.7}_{-1.9}$
$U(\psi_{20\text{phn}})$	Triplet unigrams	$75.6\%^{+2.0}_{-2.1}$	$9.3\%^{+1.1}_{-1.2}$	$16.4\%^{+1.4}_{-1.3}$	$26.2\%^{+1.3}_{-1.4}$
PPRLM	PPRLM System	$87.4\%^{+1.5}_{-1.5}$	$4.9\%^{+0.7}_{-0.6}$	$10.3\%^{+1.4}_{-1.4}$	$19.2\%^{+1.5}_{-1.4}$

Table 10
ID rate and equal error rates for fused systems

Fusion experiment	30 s Test (ID Rate)	30 s Test (EER)	10 s Test (EER)	3 s Test (EER)
$\lambda(\psi) + \lambda_{\text{pitch}}(\psi)$	$82.5\%^{+1.8}_{-1.8}$	$6.8\%^{+1.1}_{-0.8}$	$11.6\%^{+1.2}_{-1.1}$	$20.0\%^{+1.4}_{-1.2}$
$\lambda(\psi) + \lambda_{\text{pitch}}(\psi) + U(\psi_{20\text{phn}})$	$89.3\%^{+1.5}_{-1.5}$	$4.5\%^{+0.7}_{-0.6}$	$10.1\%^{+1.1}_{-0.9}$	$19.7\%^{+1.3}_{-1.4}$
$\lambda(\psi) + \lambda_{\text{pitch}}(\psi) + U(\psi_{20\text{phn}}) + PPRLM$	$92.8\%^{+1.2}_{-1.3}$	$2.6\%^{+0.6}_{-0.5}$	$7.8\%^{+1.0}_{-1.0}$	$16.8\%^{+1.3}_{-1.5}$
$\lambda(\psi_{6\text{Class}}) + \lambda_{\text{pitch}}(\psi) + U(\psi_{20\text{phn}}) + \text{PPRLM}$	$93.1\%^{+1.2}_{-1.2}$	$2.2\%^{+0.5}_{-0.5}$	$7.4\%^{+1.0}_{-0.9}$	$16.5\%^{+1.3}_{-1.2}$

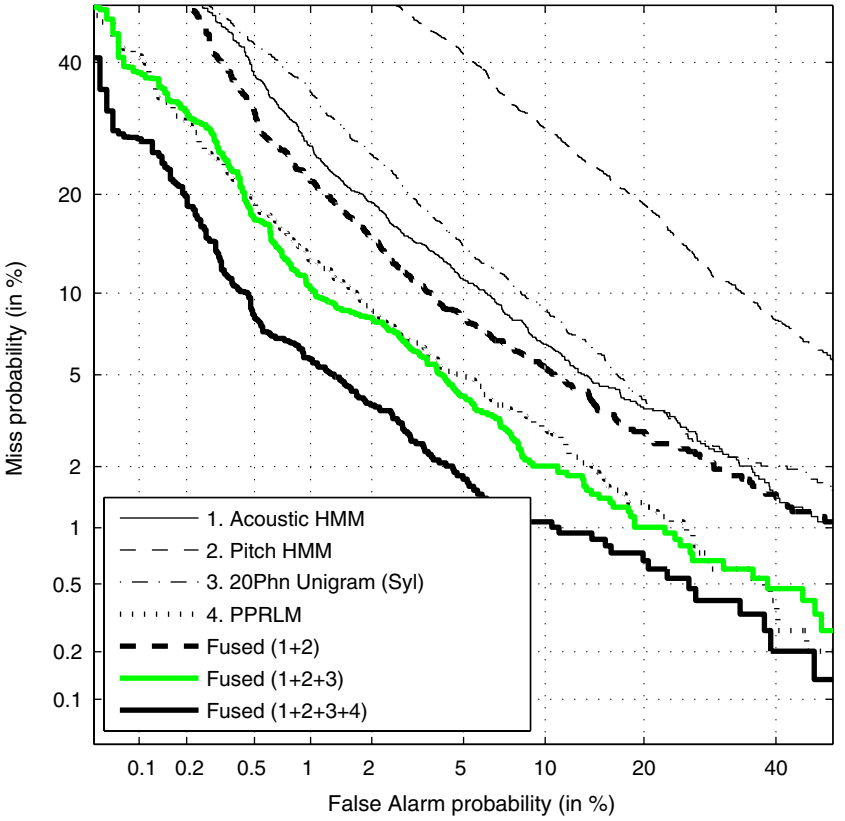

Fig. 8. DET plot showing individual and fused system performance.

Three fusion experiments were performed and the results are provided alongside statistical significance figures in Table 10. Additionally, Fig. 8 illustrates the performance of the various fusion experiments for the 30 s test. The following discussion refers to results for the 30 s test, although generally similar results were observed for the 10 and 3 s tests.

The first experiment combined the HMM acoustic system $\lambda(\psi)$, and the pitch system based around the triplet framework, $\lambda_{\text{pitch}}(\psi)$. Initial inspection of Fig. 8 indicates that the pitch system provides a minor improvement to the HMM system. However, Tables 9 and 10 reveal that the drop from 8.0% EER for the HMM system to 6.8% EER after fusing the pitch information is not statistically significant.

As shown in Section 7, the addition of the coarse 4 class unigram statistics did not provide any complementary information beyond that obtained from the PPRLM and $U(\psi_{20\text{phn}})$ systems. Accordingly, only the fine phonetic triplet unigram scores $U(\psi_{20\text{phn}})$, were combined with the previous fusion combination. This provides a statistically significant absolute improvement of 2.3%. Of further note is that this combination, which consists of scores obtained only using the syllabic framework, is comparable to the score obtained by the stand alone PPRLM system (PPRLM).

In order to examine whether the proposed framework adds anything to a stand-alone PPRLM system the three triplet based systems, $\lambda(\psi)$, $\lambda_{\text{pitch}}(\psi)$, and $U(\psi_{20\text{phn}})$ are combined with the

PPRLM system. This combination improves the performance from 4.9% EER to 2.6%, which again is significant. Finally, the systems based on the expanded 6 phone set for both the HMM acoustic system and pitch system were examined. The use of the HMM system based on 6 classes $\lambda(\psi_{6\text{Class}})$, produced an additional improvement of 0.4% taking the EER to 2.2%. However, the inclusion of the 6 Class pitch system provided no additional benefit and so results in Table 10 only includes the result for the additional fusion of $\lambda(\psi_{6\text{Class}})$.

10. Conclusions

In this study a syllable-length segmental framework was used to analyse how individual information sources contribute to overall language identification performance. The syllabic framework was achieved via a multilingual phone recognition system, which used broad phonetic classes.

Features derived to represent acoustic, prosodic and phonotactic information were then used to produce three separate models. The first series of experiments, based on modelling acoustic features, was then conducted. A baseline GMM system was compared with a GMM system which modelled the recognised segments, with both systems achieving comparable levels of performance. Attempts to improve the temporal modelling of the phone-triplets provided marginal improvements, however these were not significant, thereby reinforcing previous observations by Zissman and Singer (1995) that HMM modelling does not exploit the available temporal information.

However, the segmental framework did provide the opportunity to examine which phone-triplets were providing the most discrimination, with most of the discrimination existing in the top 8 of the 64 syllables. Of further note was that vowels and nasals appear more consistently in these syllables.

A second series of experiments examined whether complementary phonotactic information could be extracted by using n-gram statistics over both short and extended segmental lengths. Additionally, the suggestion that complementary information is also contained in both fine and coarse phonetic representations was also assessed. It was found that the use of unigrams statistics for the phone-triplets provided significant improvements when used to complement existing PPRLM systems. However no information is contained within the coarse representation of syllabic events, which cannot be gleaned from those units containing more phonetic resolution.

The framework was also used to model prosodic dynamics, achieving similar levels of performance to that proposed by Adami and Hermansky (2003), and importantly provides direction for improving this performance.

Finally, a small set of fusion experiments was conducted in order to assess the degree of complementary information contained within the acoustic, phonotactic and prosodic systems. The best performing acoustic system, based on the HMM models, was combined with the pitch system, providing a minor improvement.

The levels of performance achieved by the baseline prosodic system and that built under the syllabic framework were comparable, with results indicating that prosodic information can be used to obtain marginal improvements when combined with acoustic and phonotactic systems. However the subsequent combination of the HMM acoustic, prosodic and phone-triplet unigrams achieved similar levels of performance to the PPRLM system and importantly, the fusion of all systems resulted in an absolute improvement of 5.4% in ID rate and 2.3% absolute EER, over

the stand alone PPRLM system, for the 30 s test. Similar improvements were observed for the 10 and 3 s test.

A common theme to emerge in this study was that the judicious selection of the more discriminatory syllabic events may assist in extracting further improvements for both the individual and fused systems. Using the information made possible by the framework, the broad class phone-set was expanded from four to six, providing further improvements to the acoustic system, but not the pitch based system. This anomaly will be the subject of future investigation in future research.

Acknowledgements

The authors thank Michael Mason and Robbie Vogt for their technical advice and editorial support.

References

Adami, A., Hermansky, H., 2003. Segmentation of speech for speaker and language recognition. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Geneva, pp. 841–844.

Baker, B., Vogt, R., Mason, M., Sridharan, S., 2004. Improved phonetic and lexical speaker recognition through MAP adaptation. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 94–99.

Berkling, K.M., 1996. Automatic language identification with sequences of language independent phoneme clusters. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology.

Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers. In: Eurospeech, vol. 4, Denmark, pp. 2517–2520.

Wong, E., Martin, T., Svendsen, T., Sridharan, S., 2003. Multilingual phone clustering for recognition of spontaneous Indonesian speech utilising pronunciation modeling techniques. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Geneva, pp. 3133–3136.

Foil, J., 1986. Language identification using noisy speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 861–864.

Goodman, F., Martin, A.F., Wohlford, R.E., 1989. Improved automatic language identification in noisy speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 528–531.

Hazen, T., Zue, V., 1997. Segment-based automatic language identification. J. Acoust. Soc. Am. 101 (4), 2323–2332.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752.

Hieronymus, J., Kadambe, S., 1996. Spoken language identification using large vocabulary speech recognition. In: Proceedings of the International Conference on Spoken Language Processing, vol. 3, pp. 1780–1783.

House, A., Neuberg, E., 1977. Toward automatic identification of the language of an utterance. i. Preliminary methodological considerations. J. Acoust. Soc. Am. 62 (3), 708–713.

Hutchins, S., Thyme-Gobbel, A., 1994. The role of prosody in language identification. In: Proceedings of the 15th Annual Speech Research Symposium, pp. 76–83.

Kajarekar, S., 2002. Analysis of variabilty in speech with applications to speech and speaker recognition. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, Portland, USA.

Li, K.P., Edwards, T.J., 1980. Statistical models for automatic language identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 884–887.

Linguistic Data Consortium, 1996. CallFriend corpus. Available from: <http://www.ldc.upenn.edu/>.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), vol. 4, pp. 1895–1898.

Martin, T., Wong, E., Baker, B., Mason, M., Sridharan, S., 2004. Pitch and energy trajectory modelling in a syllable length temporal framework for language identification. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 289–296.

Massachusetts Institute of Technology Lincoln Laboratory, 2004. LNKnet Pattern Classification Software. Available from: <http://www.ll.mit.edu/IST/lnknet/>.

Mendoza, S., Gillick, L., Ito, Y., Lowe, S., Newman, M., 1996. Automatic language identification using large vocabulary continuous speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 785–788.

Muthusamy, Y., Cole, R., Oshika, B., 1992. The OGI multi-language telephone speech corpus. In: International Conference on Spoken Language Processing, pp. 895–898.

Muthusamy, Y.K., 1993. July A segmental approach to automatic language identification. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology.

Nakagawa, S., Seino, T., Ueda, Y., 1994. Spoken language identification by ergodic HMM's and its state sequences. Electron. Commun. Jpn. 77 (6), 70–79.

Navratil, J., 2001. Spoken language recognition – a step towards multilinguality in speech processing. IEEE Trans. Speech Audio Process. 9 (9), 678–685.

Parris, E., Carey, M., 1995. Language identification using multiple knowledge sources. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 3519–3522.

Reynolds, D., 1997. Comparison of background normalization methods for text-independent speaker verification. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Vol. 2, pp. 963–966.

Savic, M., Acosta, E., Gupta, S.K., 1991. An automatic language identification system. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 528–531.

Singer, E., Torres-Carrasquillo, P., Gleason, T., Campbell, W., Reynolds, D., 2003. Acoustic and discriminative approaches to automatic language identification. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Geneva, pp. 1345–1349.

Talkin, D., 1995. Speech Coding and Synthesis. Elsevier, New York.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Green, R., Reynolds, D., 2002. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: Proceedings of the International Conference on Spoken Language Processing, vol. 1, Denver, pp. 89–92.

Tucker, R., Carey, M., Parris, E., 1994. Automatic language identification using sub-word models. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 301–304.

Wong, E., Sridharan, S., 2003. Spoken language identification utilising both acoustic and phonetic information. In: International Symposium on Digital Signal Processing and Communication Systems, pp. 520–526.

Yan, Y., Barnard, E., 1995. An approach to automatic language identification based on language-dependent phone recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 3511–3514.

Yan, Y., Barnard, E., Cole, R., 1996. Development of an approach to automatic language identification based on phone recognition. Computer Speech Lang. 10, 37–54.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK Book for Version 3.2. Entropic.

Zissman, M., Singer, E., 1994. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modelling. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 305–308.

Zissman, M., Singer, E., 1995. Language identification using phoneme recognition and phonotactic language modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 3503–3506.