Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 48 (2006) 913–926
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Language identification using acoustic log-likelihoods of syllable-like units

T. Nagarajan *, H.A. Murthy

Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India

Received 29 November 2004; received in revised form 30 September 2005; accepted 20 December 2005

Abstract

Automatic spoken language identification (LID) is the task of identifying the language from a short utterance of the speech signal uttered by an unknown speaker. The most successful approach to LID uses phone recognizers of several languages in parallel [Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. 4 (1), 31–44]. The basic requirement to build a parallel phone recognition (PPR) system is segmented and labeled speech corpora. In this paper, a novel approach is proposed for the LID task which uses parallel syllable-like unit recognizers, in a frame work similar to the PPR approach in the literature. The difference is that the sub-word unit models for each of the languages to be recognized are generated in an unsupervised manner without the use of segmented and labeled speech corpora. The training data of each of the languages is first segmented into syllable-like units and language-dependent syllable-like unit inventory is created. These syllable-like units are then clustered using an incremental approach. This results in a set of syllable-like units models for each language. Using these language-dependent syllable-like unit models, language identification is performed based on accumulated acoustic log-likelihoods. Our initial results on the Oregon Graduate Institute Multi-language Telephone Speech Corpus [Muthusamy, Y.K., Cole, R.A., Oshika, B.T., 1992. The OGI multi-language telephone speech corpus. In: Proceedings of Internat. Conf. Spoken Language Process., October 1992, pp. 895–898] show that the performance is 72.3%. We further show that if only a subset of syllable-like unit models that are unique (in some sense) are considered, the performance improves to 75.9%.

Keywords: Language identification; Syllable; Incremental training

1. Introduction

Automatic spoken language identification without any knowledge about the languages to be identified is a challenging problem. In the spoken language identification task, it should be assumed that no test speaker's spectral or any other type of information is present in the training set. In that, the comparison between the test utterance and the reference models of the languages is from unconstrained utterances of two different speakers (Li, 1994). Therefore, the differences between two utterances encompass text differences, speaker

* Corresponding author. Address: University of Quebec, INRS-EMT, 800, de la Gauchetiere ouest, Bureau 6900, Montreal, Quebec, Canada H5A 1K6. Fax: +1 514 875 0344.
E-mail address: raju@emt.inrs.ca (T. Nagarajan).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.12.003

differences, environment differences, and language differences. The main problem is how to extract the language differences apart from text, speaker, and environment differences for a reliable spoken language identification system.

The main features of an ideal spoken language identification system are:

- The computation time requirement to determine the identity of a test utterance must be small.
- The performance degradation must be graceful as the length of the test utterance is reduced.
- The system should not be biased towards any language or a group of languages.
- The system should not be complex, in the sense that,
  - amount of language specific information required for developing the system should be small,
  - including a new language into the existing system should be easy.
- The system should tolerate,
  - channel and environment variations,
  - noisy/low SNR speech signals,
  - accent variations.

Humans are the best LID systems in the world today. Just by hearing one or two seconds of speech of a **familiar language**, they can easily identify the language. The sources of information used by humans to identify the language are several. Speech in a language is a sequence of phones/sound units and the differences among the languages can be at several levels. Hierarchically, these levels are Frame level (10–30 ms), phone level, consonant-vowel (CV unit) level, syllable level, word level, and phrase level. The possible differences among different languages at these levels are the inventory, the frequency of occurrence of different units in each set, the sequence of units (phonotactics) and their frequencies of occurrence, the acoustic signatures, the duration of the same sound unit in different languages, and the intonation patterns of units at higher levels. The performance of any LID system depends on the amount of information and the reliability of information extracted from the speech signal and how efficiently it is incorporated into the system.

Existing spoken language identification systems can be broadly classified into two groups, namely, **explicit and implicit LID systems**. The LID systems that require speech recognizers of one or several languages, in other words, the systems that require a segmented and labeled speech corpus are termed here as **explicit LID systems**. The language identification systems which do not require phone recognizers (or rather segmented and labeled speech data) are termed here as **implicit LID systems**. In other words, these systems require only the raw speech data along with the true identity of the language spoken (Zissman, 1996). The language models or the language-specific information are derived only from the raw speech data. In the literature both types of systems have received significant attention.

A number of researchers have used phone recognizers (either language-dependent or language-independent) as front-end for language identification (Lamel and Gauvain, 1994; Berkling et al., 1994; Hazen and Zue, 1994; Kadambe and Hieronymus, 1995; Yan and Barnard, 1995; Navratil and Zuhlke, 1997). The most successful approach to LID (in terms of performance) uses phone recognizers of several languages in parallel (Zissman, 1996). In (Zissman, 1996), it is shown that even with one language phone recognizer, a language identification system can be built. But the analysis in (Zissman, 1996) also indicates that the performance of the system considerably improves in proportion to the number of front-end phone recognizers. The basic requirement for building a parallel phone recognition (PPR) system is a segmented and labeled speech corpus. Building segmented and labeled speech corpora for all the languages to be recognized, is both time consuming and expensive, requiring trained human annotators and substantial amount of supervision (Greenberg, 1999). Further, in (Singer et al., 2003), a GMM and an SVM based implicit LID systems are shown to perform better than the conventional explicit LID systems. Therefore, the unavailability of segmented and labeled speech corpus, and the recent developments in the implicit LID systems, make the implicit LID systems more attractive.

In (Jayaram et al., 2003; Ramasubramanian et al., 2003), a parallel sub-word recognition system for the LID task is proposed, in a framework similar to the parallel phone recognition (PPR) approach in the literature (Zissman, 1996). The difference is that this approach does not require segmented and labeled speech corpora. Since most of the phonemes among different languages are common, the source of information that may be used for LID is the variation in the frequency of occurrence of the same phoneme in different languages, and the variation in

the acoustic realization of the same phoneme in different languages. Only very few phonemes are unique for a particular language. If a longer sound unit, say syllable-like unit is used, then the number of unique syllable-like units in any language is very high, which may be a potential information for discriminating languages. Li (1994) proposed a system which is based on features extracted at the syllable level. In this system, the syllable nuclei (vowels) for each speech utterance are located automatically. Next, feature vectors containing spectral information are computed for regions near the syllable nuclei. Each of these vectors consists of spectral sub-vectors computed on neighboring frames of speech data. Rather than collecting and modeling these vectors over all training speech, Li keeps separate collections of feature vectors for each training speaker. During testing, syllable nuclei of the test utterance are located and feature vector extraction is performed. Each speaker-dependent set of training feature vectors is compared to feature vectors of the test utterance, and most similar speaker-dependent set of training vectors is found.

One of the major reasons for considering the syllable as a basic unit for speech recognition systems is its better representational and durational stability relative to the phoneme (Wu et al., 1998). The syllable was proposed as a unit for ASR as early as 1975 (Fujimura, 1975), in which irregularities in phonetic manifestations of phonemes were discussed. It was argued that the syllable will serve as an effective minimal unit in the time-domain. In (Prasad, 2003), it is demonstrated that segmentation at syllable-like units followed by isolated style recognition of continuous speech performs well.

Many languages of the world possess a relatively simple syllable structure consisting of several canonical forms (Greenberg, 1999). Most of the syllables in such languages contain just two phonetic segments, typically of **CV** type (for example, Japanese). The remaining syllabic forms are generally of **V** or **VC** variety. In contrast, English and German possess a more highly heterogeneous syllable structure. In such forms, the onset and/or coda constituents often contain two or more consonants. But a salient property shared in common by stress and syllable-timed languages is the preference for **CV** syllabic forms in *spontaneous speech*. Nearly half of the forms in English and over 70% of the syllables in Japanese are of this variety. There is also a substantial proportion of **CVC** syllables in spontaneous speech of both the languages (Greenberg, 1999). This shows that even for the languages which are not syllable-timed, the syllable can be defined using a simple structure. Further, the definition of syllable in terms of short-term energy function is suitable for almost all the languages, in the case of spontaneous speech.

In this paper, a novel approach is proposed for the LID task which uses parallel syllable-like unit recognizers (Nagarajan and Murthy, 2004), in a framework similar to PPR approach in the literature with one significant difference. The difference is that the sub-word unit models (syllable-like unit models) for each of the languages to be recognized are generated in an unsupervised manner without the use of segmented and labeled speech corpora.

The basic requirement for building syllable-like unit recognizers for all the languages to be identified, is an efficient segmentation algorithm. Earlier, an algorithm (Prasad et al., 2004) was proposed, which segments the speech signal into syllable-like units. Recently, several refinements (Nagarajan et al., 2003) have been made to improve the segmentation performance of the baseline algorithm (Prasad et al., 2004). Using this algorithm (Nagarajan et al., 2003) each language training utterances are first segmented into syllable-like units. Similar syllable segments are then grouped together and syllable models are trained incrementally. These language-dependent syllable models are then used for identifying the language of the unknown test utterances.

The rest of the paper is organized as follows. In Section 2, the speech corpora used for this study is mentioned. In Section 3, the segmentation approach used to segment the speech signal into syllable-like units is briefly described. Section 4 describes the unsupervised and incremental clustering procedure which is used to cluster similar syllable-like units. In Section 5, different methods used to identify the language of the unknown utterance are described in detail. The performance of these LID systems are analyzed in Section 6.

2. Speech corpus

The Oregon Graduate Institute Multi-language Telephone Speech (OGI_MLTS) Corpus (Muthusamy et al., 1992), which is designed specifically for LID research, is used for both training and testing. This corpus currently consists of spontaneous utterances in 11 languages: English (En), Farsi (Fa), French (Fr), German (Ge), Hindi (Hi), Japanese (Ja), Korean (Ko), Mandarin (Ma), Spanish (Sp),

Tamil (Ta) and Vietnamese (Vi). The utterances were produced by ~90 male and ~40 female, in each language over real telephone lines. In our work, presently all the 11 languages are used. To maintain the homogeneity in training and testing across languages, for each language first 35 male speakers' and first five female speakers' 45 s utterances are used for training and next 18 male speakers' and next two female speakers' 45 s utterances are used for testing. The rest of the data is used as development set. All the training and test set speakers are different.

3. Segmentation of speech into syllable-like units

Researchers have tried different ways of segmenting the speech signal either at the phoneme level or at the syllable level (Mermelstein, 1975; Schmidbauer, 1987; Nakagawa and Hashimoto, 1988; Noetzel, 1991;Shastri et al., 1999), with or without the use of phonetic transcription. These segmentation methods can further be classified into two categories, namely, time-domain based methods, where short-term energy function, zero-crossing rate, etc. are used and frequency-domain based methods, where short-term spectral features are used.

Earlier a method was proposed (Prasad et al., 2004) for segmenting the acoustic signal into syllable-like units,[1] in which a minimum phase signal is derived from the symmetrized and inverted[2] Short-term energy (STE) function of the speech signal as if it were a magnitude spectrum. It is observed that the group delay function of this minimum phase signal is a better representative of the short-term energy function to perform segmentation. Later, several refinements have been made to improve the performance of the baseline segmentation algorithm (Nagarajan et al., 2003).

In this approach, for both training and testing, only 45 s utterances of OGI_MLTS corpus are used. The segmentation approach discussed in (Nagarajan et al., 2003) can segment the speech files of any duration without significant degradation in performance. Here, each 45 s utterance is given as a whole to the segmentation algorithm. The segmentation performance is quite satisfactory (see Fig. 1), except one noticeable problem (see the marked region in Fig. 1). Since the group delay function exhibits an additive property, the influence of the neighboring poles and zeroes (syllable boundaries and the center of the syllable nuclei) is very small. But in some cases, it cannot be neglected.

Merging of two syllables, or in other words, a segment boundary may be missed because of the influence of the rest of the peaks, and valleys in the system. The influence may be strong when the number of peaks, and valleys in the system is very high, i.e., when the number of syllable segments in the given speech signal is very high. Durational analysis (Greenberg, 1999) made on Switchboard corpus (Godfrey et al., 1992) shows that the mean duration of the syllables is ≈200 ms. When a whole 45 s utterance is given for segmentation, if the duration of any of the segment is found to be more than 200 ms (see Fig. 1), that particular segment alone is extracted from the original speech signal (see Fig. 2), and the segmentation algorithm is applied once again on the extracted segment. If a positive peak is detected in between, that segment is split into two (see Fig. 3).

Using this approach, all the training speech data of each language are segmented into syllable-like units, which gives K_{L_i} syllable-like units, $s_1, s_2, \ldots, s_{K_{L_i}}$ ($K_{L_i} \approx 6000$) for the language L_i. These syllable-like units are used during the training process. The training process is similar to the conventional clustering technique but instead of clustering the feature vectors at frame level (10–30 ms), it is done at syllable-like unit level (75–325 ms), using the method described in the following Section.

4. Unsupervised and incremental clustering

The main objective of this work it to derive a minimal set of syllable-like unit models for each language, to carry out the language identification task. Here, the Hidden Markov modeling (HMM) technique is used to model the automatically segmented syllable-like units, and to reduce the number of syllable-like unit models of each language. To derive sub-word unit models, the conventional batch training technique can be used in which all the training examples, which belong to a particular class, are given at once. For this technique, the basic requirement is a segmented and labeled speech corpus. But, in the present work, since the segments of training speech data do not have any identity, batch training cannot be used for deriving syllable-like unit models. In order to derive a representative set of

[1] Segment of speech in between two consecutive energy valleys is defined as syllable-like unit.

[2] Since the STE function is inverted, the peaks (poles) in the STE function become valleys (zeroes) and vice-versa.

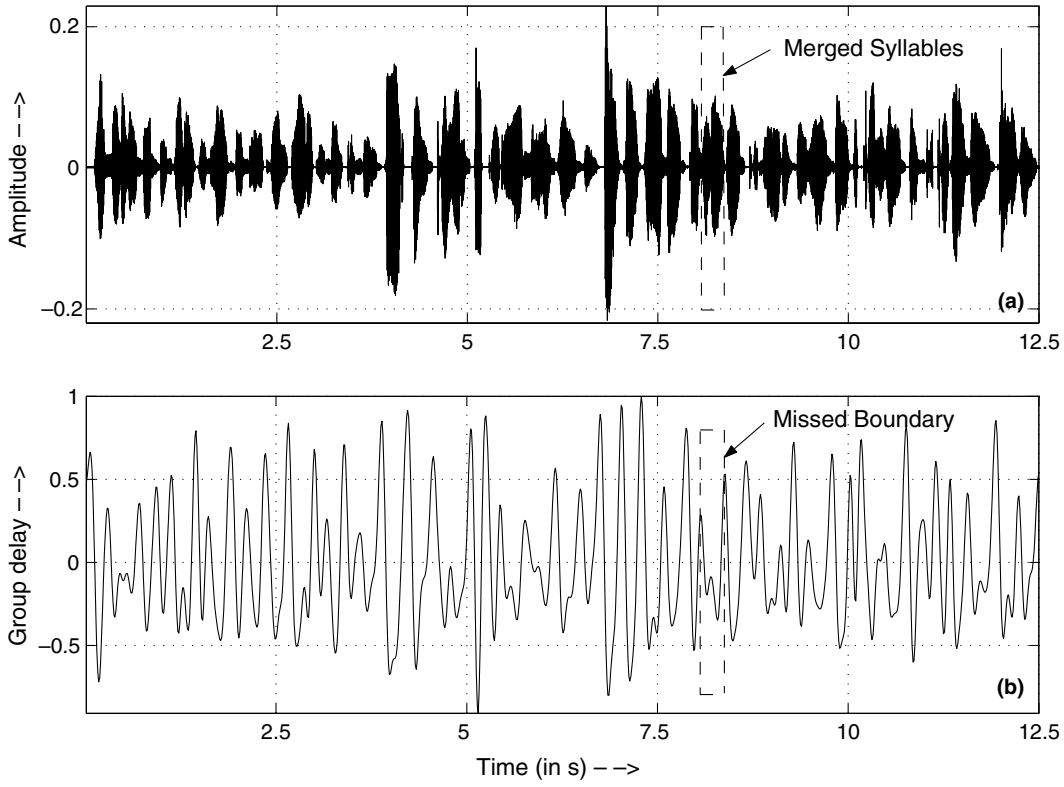

Fig. 1. (a) 12.5 s speech signal extracted from OGI Tamil data. (b) The group delay function derived from the energy contour.

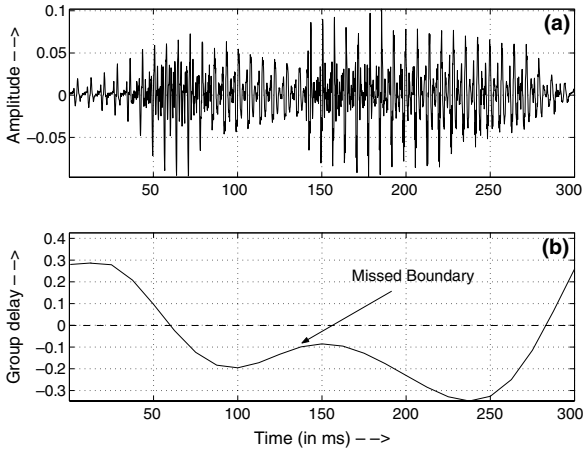

Fig. 2. (a) Expanded segment (marked region in Fig. 1). (b) Expanded group delay function (marked region in Fig. 1).

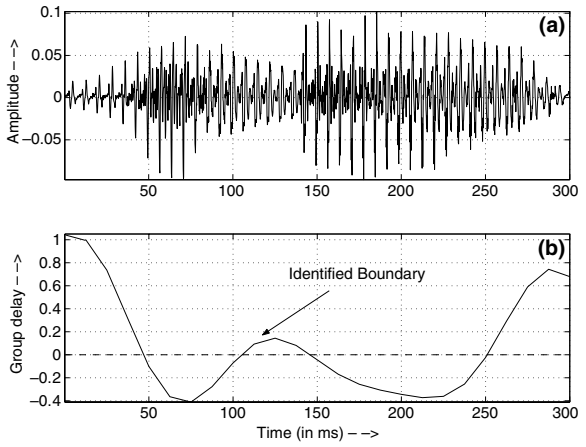

Fig. 3. (a) Extracted segment of speech. (b) New group delay function of the extracted segment alone.

syllable-like unit models, a novel clustering technique, which is referred as **unsupervised and incremental clustering**, is proposed that automatically groups similar syllable-like units. The steps for deriving the syllable-like unit models using this technique are described below (refer Fig. 4).

4.1. Initial cluster selection

For any iterative training process, the assumed initial condition is crucial for the speed of convergence. After all the syllable-like units have been obtained, the first task is to select some unique

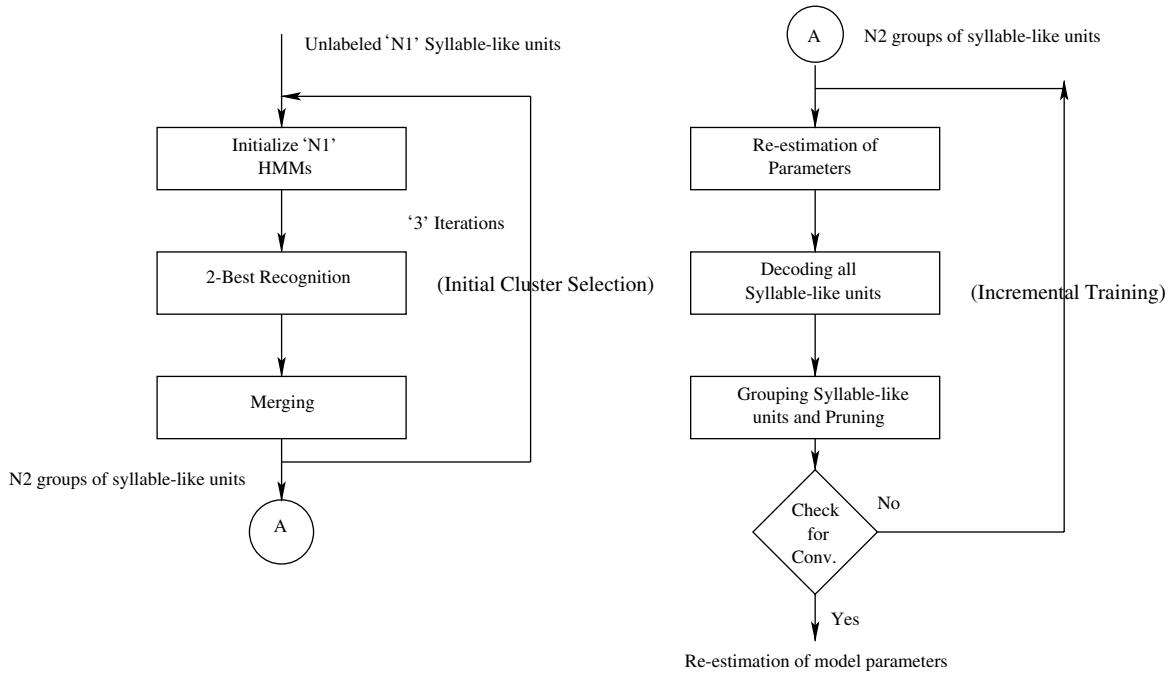

Fig. 4. Flow chart: unsupervised and incremental HMM.

syllable-like units or groups of unique syllable-like units for training. The initial groups of syllable-like units should be carefully chosen to ensure fast convergence. At the initial stage itself, if the selected group of syllable-like units are unique, the convergence may be accelerated during iterative training. For selecting such initial clusters, the following procedure is adopted.

(1) From the K_{L_i} syllable-like units of the language L_i, a subset ($N1$) of syllable-like units,[3] $s_1, s_2, \ldots, s_{N1}$, where $N1 < K_{L_i}$, are taken randomly for initialization (here, $N1 = 2000$).
(2) Features (13-dimensional MFCC + 13 delta + 13 acceleration coefficients, after cepstral mean subtraction) are extracted from these $N1$ syllable-like units with multiple resolutions (i.e., with different window sizes). The magnitude spectra and the resultant MFCC features extracted from the same speech signal with different window sizes are considerably different due to different frequency resolutions (Rabiner and Schafer, 1978). This process artificially introduces more examples for the same class (which are derived from a single example) and in turn it ensures a reasonable variance for each Gaussian mixture in the models. To show the effect of multi-resolution feature extraction, an experiment is conducted as explained below. For this experiment, 200 automatically segmented syllable-like units from Tamil speech data of OGI_MLTS corpus are considered for training and the same set is used for testing. During training, in the case of MRFE, the features are extracted with five different window sizes (12, 14, 16, 18, and 20 ms) with a fixed frame-shift of 10 ms, and for the single-resolution feature extraction (SRFE), the features are extracted with a single window size (20 ms). Hidden Markov models (HMM) for the 200 syllable-like units are initialized in both SRFE and MRFE separately. During testing, from the same 200 syllable-like units, the features are extracted with nine different window sizes (13, 15, 17, 19, 20, 21, 23, 25, and 27 ms) which are not used during training and tested against both (SRFE and MRFE) the set of syllable models. Here, the feature set extracted with each window size is considered as a separate test set (results in nine different test sets) and the performance for all the test sets are given in the

[3] A subset of K_{L_i} alone is considered in order to reduce the computational complexity. In fact, all K_{L_i} segments can be used.

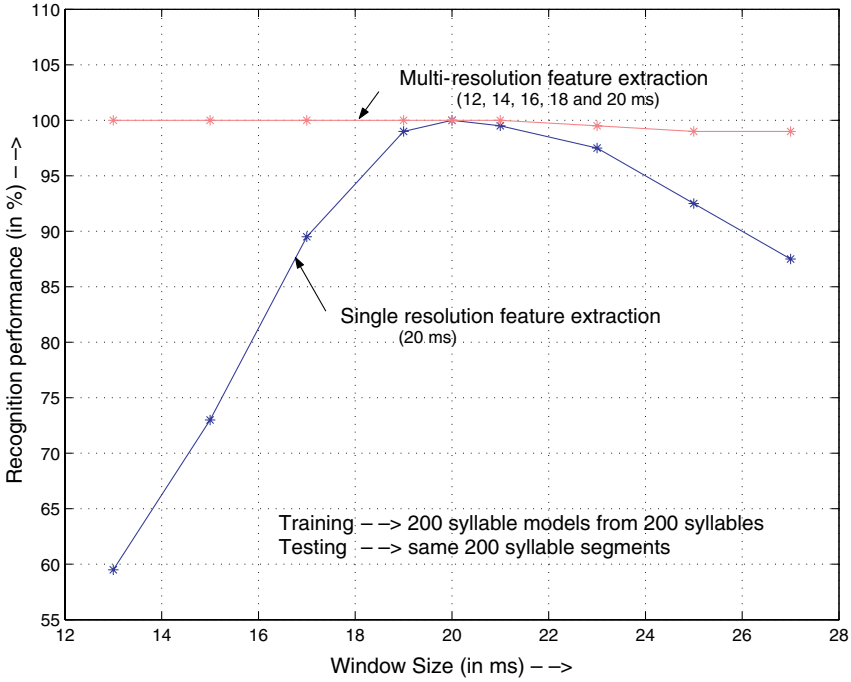

Fig. 5. Comparison between multiple resolution and single resolution feature extraction.

Fig. 5. This performance comparison shows that, models generated with MRFE features can handle variations (see the performance for 21, 23, 25, and 27 ms features).

(3) $N1$ *Hidden Markov Models* $(\lambda_1, \lambda_2, \ldots, \lambda_{N1})$ are initialized with five states and three gaussian mixtures/state.

(4) Using the Viterbi decoding process, the same $N1$ syllable-like segments are decoded using 2-best criteria, resulting in $N1$ pairs of syllable-like units $(P_1, P_2, \ldots, P_{N1})$.

$$P_j = \left[\arg \max^{1}_{1 \leqslant i \leqslant N1} p(O|\lambda_i), \quad \arg \max^{2}_{1 \leqslant i \leqslant N1} p(O|\lambda_i)\right] \tag{1}$$

where

- P_j is the jth pair of syllable-like units (where $1 \leqslant j \leqslant N_1$),
- $p(O|\lambda_i)$ is the probability of the observation sequence O $(o_1 o_2 \ldots o_n)$ for the given model λ_i,
- $\max^1$ and $\max^2$ denote the 1-best and 2-best results, respectively.

Interestingly, in each of the $N1$ cases, the first-best syllable is the same syllable which is used for the corresponding model initialization, and the second-best is another syllable which sounds identical/similar to the first-best syllable.

(5) Among $N1$ pairs $(P_1, P_2, \ldots, P_{N1})$, if a syllable-like unit is found to be repeated in more than one pair, the other pairs are removed and the number of models is thus pruned.

(6) New models are created with these reduced number of pairs. These new models will have the identity of the syllable-like units from which they are generated.

(7) Steps 4–6 are repeated $\boldsymbol{m}$ times (here, $\boldsymbol{m} = 3$). After $\boldsymbol{m}$ iterations, each cluster will have $2^{\boldsymbol{m}}$ syllable-like units grouped together.

This initial cluster selection procedure leads to $N2$ clusters $(C_1, C_2, \ldots, C_{N2})$, which are unique, and each cluster is expected to have similar syllable-like units. Since we start with single syllable-like unit, the model parameters are only initialized and not re-estimated/refined. The sub-sequent Section describes an incremental training procedure, where the model parameters are re-estimated and tuned at every iteration.

4.2. Incremental training procedure

After selecting the initial clusters $(C_1, C_2, \ldots, C_{N2})$, where the models are only initialized, the parameters of the models of each of the clusters are re-estimated using Baum–Welch

re-estimation procedure. The steps followed for this incremental training are given below:

(1) The model parameters of the initial clusters $(C_1, C_2, \ldots, C_{N2})$ derived from the previous step are re-estimated using Baum–Welch re-estimation. Each model is a 5 state 3 Gaussian mixtures/state HMM.
(2) The new models are used to decode all the syllable-like units $(s_1, s_2, \ldots, s_{K_{Li}})$ using Viterbi decoding.
(3) Clustering is done based on the decoded sequence.
(4) If a particular cluster is found to have less than ϵ (here, $\epsilon = 3$) syllable-like units, that cluster is removed and number of models is reduced by one.
(5) Steps 1–4 are repeated until convergence is met. This training procedure is referred to as **incremental training**. It is considered incremental because the HMM parameters are adjusted before all the training data corresponding to each of the models, have been considered. This training strategy must be contrasted to conventional batch training where the models are updated only after all the data in the training set, for each of the models, are processed. In this incremental procedure, the number of syllable segments assigned to each of the clusters after each iteration is different. This is essentially due to migration of syllable segments between clusters.

The convergence criteria followed in this approach is explained below.

4.3. Convergence criteria

In each iteration, as the model parameters are re-estimated and the syllable-like units are re-clustered, the number of syllable-like units which migrate from one cluster to another is expected to reduce from iteration to iteration. The convergence criteria followed for the incremental training is based on 'number of migrations between clusters'. The convergence is said to be met if the number of migrations between clusters reaches zero. When this condition is met, the incremental training procedure terminates. Invariably, in all the cases, the number of migrations are found to be zero after approximately 10 iterations. This incremental training process produces M_{L_i} syllable-like unit clusters $(C_1, C_2, \ldots, C_{M_{L_i}})$, and in turn M_{L_i} syllable-like unit models $(\lambda_1, \lambda_2, \ldots, \lambda_{M_{L_i}})$. The important point to note here is that, the number of syllable-like unit models for each language is not decided a priori. Among different languages, it varies between 320 and 400 models.

For each of the languages, the above-mentioned process is done separately and the language-dependent syllable-like unit models are trained. Here, the entire training process is unsupervised and so these clusters/syllable-like unit models do not have any identity. Using these language-dependent syllable-like unit models, the language identification task is carried out in different ways as explained in the sub-sequent section.

5. Language identification (LID) systems

One of the important language identification cues that can be used in parallel sub-word unit recognition based systems is the *n*-gram statistics. Even if the speech data used during training is limited, the *n*-gram statistics can very well be derived from the digital text and used for the language identification task. But, if the training process is unsupervised and if the sub-word unit models do not have any identity, *n*-gram statistics derived from the digital text cannot be of any use. An immediate alternative for this is to derive implicit *n*-gram statistics using the final clusters and their relative distributions. But, when the amount of training data is limited and the size of the sub-word unit considered is relatively large, the derived *n*-gram statistics will not contain any useful information about the language. In the present work, the number of syllable-like segments available for each language is only around 6000 and the number of automatically derived syllable-like models per language is around 350 (on an average). The number of syllable-like units clustered to each of the syllable-like models vary in between 10 and 50. Experiments were conducted to check the performance of language identification task using *n*-gram statistics and the results are found to be very poor. Even though the above-described problems make the task difficult, there are some potential features which still can be used for performing language identification.

- Recognizing a large unit like the syllable-like unit is equivalent to recognizing a trigram/bigram at the phoneme level.

- As the size of the sub-word unit is big, the number of units which are unique to a particular language is likely to be large.
- Whatever be the size of the unit, the variation in the acoustic realization of that unit among different languages can still be used as a potential candidate for the language identification task.

Using these cues, the language identification task is performed in different ways at the syllable-like unit level as described below.

5.1. LID system using accumulated acoustic likelihood (AAL)

For each language L_i, (where $i = 1, 2, \ldots, N$), a language-dependent syllable-like unit inventory is created using the segmentation algorithm described in the Section 3. Using these syllable-like unit inventories syllable-like unit models are created for each language L_i separately using the unsupervised and incremental procedure described in Section 4. The number of syllable-like unit models are not decided a priori. The clustering process automatically derives M_{L_i} syllable-like unit models $(\lambda_1, \lambda_2, \ldots, \lambda_{M_{L_i}})$ for the language L_i. During testing, each of the 45 s test utterances is segmented into syllable-like units. This results in K_T syllable-like segments, say $s_1, s_2, \ldots, s_{K_T}$. Segmenting the speech data during testing can be justified as given below:

- During training, the syllable-like unit models are generated using automatically segmented speech data.
- In (Prasad, 2003), it is demonstrated that a simple isolated-style recognition system with high recognition performance can be achieved for continuous speech recognition, if the segmentation is done a priori at syllable-like unit boundaries. The advantage of such an implementation is that there are no intensive dynamic programming based computations and mainly the effect of errors are localized (Prasad, 2003).
- As described earlier, since the n-gram statistics at syllable-like unit level does not give any language specific information, an assumption can be made that each of the syllable-like unit is generated by an independent random process (even though it is not true).

After segmenting the test utterances into syllable-like units, the individual syllable-like unit can be recognized in isolation. Using the Bayes theorem, the a posteriori probability of the individual syllable-like unit s_k, for the given language L_i, can be estimated as given below:

$$p(\lambda_m|s_k, L_i) = \frac{p(s_k|\lambda_m, L_i)P(\lambda_m|L_i)}{\sum_{m=1}^{M_{L_i}}[p(s_k|\lambda_m, L_i)P(\lambda_m|L_i)]} \tag{2}$$

Here, the a priori probability $P(\lambda_m|L_i)$ of all the syllable-like unit models for the given language L_i are assumed to be same. Normalizing the acoustic likelihoods using the evidence part (the denominator part of the RHS of Eq. (2)) may be important across languages. In the present work, such normalization is not carried out due to the following reasons: (a) as mentioned in Section 4.3, the number of models vary among languages, and (b) calculation of this evidence part is computationally expensive especially when the number of models is very high. Using these constraints, the Eq. (2) can be written as

$$p(\lambda_m|s_k, L_i) \approx p(s_k|\lambda_m, L_i) \tag{3}$$

For the given syllable-like unit s_k and for the given language L_i, if we maximize the acoustic likelihood using the syllable-like unit models of the language L_i, Eq. (3) can be written as,

$$\max_{m=1,2,\ldots,M_{L_i}} p(\lambda_m|s_k, L_i) \approx \max_{m=1,2,\ldots,M_{L_i}} p(s_k|\lambda_m, L_i) \tag{4}$$

If the log-likelihood of the acoustics is considered, then,

$$\max_{m=1,2,\ldots,M_{L_i}} \log p(\lambda_m|s_k, L_i) \approx \max_{m=1,2,\ldots,M_{L_i}} \log p(s_k|\lambda_m, L_i) \tag{5}$$

Eq. (5) gives the maximized acoustic log-likelihood of the given syllable-like unit s_k for the given language L_i. But, the test utterance contains k syllable-like units. Since, an assumption is made that each syllable-like unit of the speech utterance is generated by an independent random process, the log-likelihoods of the individual syllable-like units in the test utterance can be added and resultant accumulated log-likelihood can be written as

$$\sum_{k=1}^{K_T} \max_{m=1,2,\ldots,M_{L_i}} \log p(\lambda_m|s_k, L_i) \approx \sum_{k=1}^{K_T} \max_{m=1,2,\ldots,M_{L_i}} \log p(s_k|\lambda_m, L_i) \tag{6}$$

The language which maximizes the accumulated acoustic log-likelihood is declared as the language (l^*) of the test utterance.

Table 1
Language-wise performance of the LID systems using accumulated acoustic likelihood (AAL)

Performance for 45 s test utterances												
Language	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi	Average
Perf. in %	95	90	95	80	70	65	45	35	80	65	75	**72.27**

Table 2
Confusion matrix (LID system using AAL)

	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
English	19	0	0	1	0	0	0	0	0	0	0
Farsi	0	18	2	0	0	0	0	0	0	0	0
French	1	0	19	0	0	0	0	0	0	0	0
German	1	0	2	16	0	1	0	0	0	0	0
Hindi	1	0	1	2	14	1	0	0	1	0	1
Japanese	0	0	3	1	0	13	0	0	2	0	1
Korean	0	2	4	0	0	4	9	0	1	0	0
Mandarin	2	0	6	2	0	2	0	7	0	0	1
Spanish	2	0	2	0	0	0	0	0	16	0	1
Tamil	0	1	0	0	1	1	0	0	4	13	0
Vietnamese	1	0	2	0	0	0	0	0	0	0	17

$$l^* = \arg \max_{i=1,2,\ldots,N} \left[\sum_{k=1}^{K_T} \max_{m=1,2,\ldots,M_{L_i}} \log p(\lambda_m | s_k, L_i) \right] \quad (7)$$

Using Eq. (7), the language of all the test utterances are determined. The 1-best performance of the LID using the AAL method is given in Table 1. The average performance is severely affected by the poor performance for the languages **Korean and Mandarin**.

The observations made on the confusion matrix (see Table 2) and the False acceptance and False rejection analysis (see Table 3) show that the system is biased towards some of the languages. The specific observations are given below:

Table 3
False acceptance (FA) and false rejection (FR) (%)

	False rejection	False acceptance
En	5	4.0
Fa	10	1.6
Fr	5	12.2
Ge	20	3.3
Hi	30	0.5
Ja	35	5.0
Ko	55	0.0
Ma	65	0.5
Sp	20	4.4
Ta	35	0.0
Vi	25	1.6

- French (third column in the Table 2 and third row in the Table 3) is severely biased.
- Korean and Mandarin are severely confused with the language Japanese and French (see seventh and eighth row in the confusion matrix).
- Tamil is severely confused with Spanish (see tenth row in the confusion matrix).
- Even though the performance for the languages Farsi and Vietnam are considerably good, those languages are not biased and other languages are not confused with these two.

Removing the bias, as described in (Ramasubramanian et al., 2003) can be taken up as future research.

An analysis has been made on the performance of this LID system with different durations of test signals (see Fig. 6). It shows that even for 6 s test utterances, the performance is around 50%.

5.2. LID system using voting

This method is also based on acoustic log-likelihood only. As explained in the previous method, here also for each of the syllable-like units in the test utterance, the acoustic log-likelihood (refer Eq. (5)) of each language is found. But, instead of taking the accumulated acoustic log-likelihood for making language decision, the decision is made for each of the syllable-like units separately as given below.

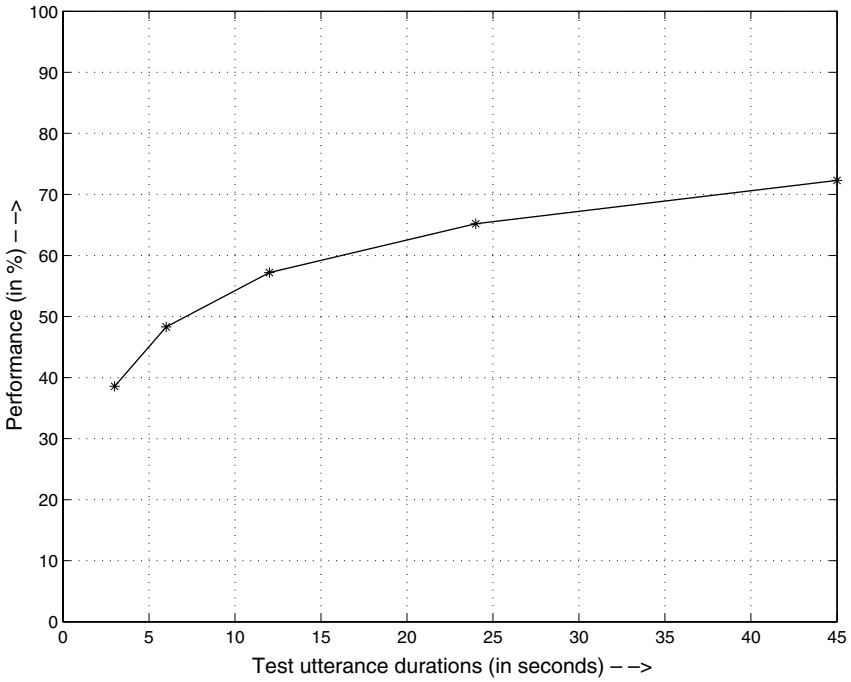

Fig. 6. Performance (AAL method) for different test utterance durations.

Eq. (5) is repeated here for reference

$$\max_{m=1,2,\ldots,M_{L_i}} \log p(\lambda_m|s_k,L_i) \approx \max_{m=1,2,\ldots,M_{L_i}} \log p(s_k|\lambda_m,L_i) \quad (8)$$

Using this maximized log-likelihoods of acoustics of each of syllable-like units separately, the language identification is carried out as below:

$$l^*_{s_k} = arg_{i=1,2,\ldots,N}\left[\max_{m=1,2,\ldots,M_{L_i}} \log p(\lambda_m|s_k,L_i)\right] \quad (9)$$

where $l^*_{s_k}$ is the language of the syllable-like unit, s_k.

The final decision is made based on number of syllable-like units which are in favour of each language L_i. The language which gets maximum number of syllable-like units in favour of it, is declared as the language (l^*) of the test utterance. The 1-best language identification performance of this system is given in Table 4.

Invariably, for all the languages the performance is either poor or at most equal to the performance of the system which used AAL. This shows that the signature of the language is not well captured in the number of syllable-like units which are in favour of it, but on how well the test syllable-like units fits in to one language. Further, the results of the system which uses AAL and this system do not complement each other, as expected.

5.3. LID system using unique syllable-like unit models

As mentioned in Section 1, if the size of the speech unit considered for recognition is large, the number of unique speech units which belong to a particular language will also be large. In the present work, since syllable-like unit is considered for recognition, we can expect many unique syllable-like

Table 4
Language-wise performance of the LID system using voting

Performance for 45 s test utterances												
Language	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi	Average
Perf. in %	80	80	90	60	55	65	40	25	65	70	60	**62.7**

Table 5
Language-wise performance of the LID systems using accumulated acoustic likelihood, US alone, and the system using AAL and unique syllable (US) segments

Method	1-best performance in %											
	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi	Average
AAL	95	90	95	80	70	65	45	35	80	65	75	**72.27**
US	80	65	85	65	80	55	55	40	70	55	60	**64.5**
AAL + US	80	80	90	85	85	90	60	40	85	65	80	**75.9**

units for each of the languages considered for identification. Since the system described in this work is totally unsupervised and since the syllable-like unit models do not have any identity, there is no straight forward method to find out the unique units of any language. But, if the syllable-like unit models ($\lambda_m^{L_i}$) of one language (say L_i) is used to recognize the syllable-like units ($s_k^{L_j}$) of the other language (say L_j), one can expect that no test units in $s_k^{L_j}$ will be recognized as one of the unique models (if any) in $\lambda_m^{L_i}$.

For each pair of the languages (say L_i, L_j, where $i \neq j$), the unique syllable models of L_i with respect to L_j and vice versa are found as explained below:

(1) Consider the syllable-like unit models ($\lambda_m^{L_i}, \lambda_m^{L_j}$) and the automatically segmented syllable-like segments ($s_k^{L_i}, s_k^{L_j}$) of the pair of languages L_i and L_j.
(2) Recognize all of the $s_k^{L_j}$ using the models of the language L_i ($\lambda_m^{L_i}$). Here, one can notice that all of the $s_k^{L_j}$ will be grouped only to a subset (say A) of the $\lambda_m^{L_i}$.
(3) Consider the rest of the models (i.e., $\lambda_m^{L_i} \cap A$) as the unique syllable models of L_i.
(4) Repeat the steps (2) and (3) by considering the $s_k^{L_i}$ and the $\lambda_m^{L_j}$ to find out the unique syllable models of the L_j.

The whole procedure should be carried out for all the pairs of the languages under consideration. During testing, each test utterance is first segmented into syllable-like units. These syllable-like units are then decoded using the syllable-like unit models of each pair of languages, say L_i and L_j. After decoding, for each of the languages in the pair, the number of unique syllable-like units are found. The language which gets maximum number of unique syllable-like units is noted as the winner for the test utterance, in that pair of languages. The 1-best performance of this system is given in Table 5. For this system also, the performance is not better than the system which uses AAL, but the results in many cases are complementary to that of AAL system.

Since the results of the above-mentioned methods are complementary in many cases, it is decided to go for **one followed by another** approach. For the 2-best languages declared by the LID system using AAL, the LID system using unique syllable-like unit models approach is used and observed a considerable improvement in the performance (refer Table 5).

6. Discussion

A thorough analysis of the errors was performed on the above-described LID methods. It is noticed that the error in identifying the languages correctly is either because of the low-quality of the speech signal or accent variation. In particular, for the language Tamil, majority of the failure cases belong to the category of utterances of Srilankan Tamils. Even though, the performance of the syllable-like unit based LID system is reasonably good, some languages are strongly biased.

The performance of the final system which uses accumulated acoustic likelihood and the unique syllable-like unit models is better than the performance of the existing implicit language identification systems (≈55% for GMM based system) but still inferior to that of the explicit language identification systems (≈89% for parallel PRLM based system (Zissman, 1996)) which uses several language phone recognizers as a front-end. A detailed performance comparison made on different LID systems can be found in (Muthusamy et al., 1994). In the syllable-like unit based LID system also, if gender-dependent language models are used (which is presently not tried), a significant improvement in the performance can be expected.

7. Conclusion

In this paper, a novel approach is proposed for spoken language identification which uses the features derived from syllable-like units. Even though the frame work used here is similar to PPR approach given in literature, the main difference is that this approach does not require segmented and labeled speech corpus of any language. Using the automatically segmented speech data, it is shown that syllable-like unit models can be generated without any supervision. For this, a clustering technique is proposed which clusters the syllable-like units and derive a set of language-dependent syllable-like unit models. With the help of these syllable-like unit models, the language identification task is carried out in different ways. It is demonstrated that using the acoustic likelihoods of the syllable-like units alone, a reasonable language identification accuracy can be achieved. Further, it is shown that unique syllable-like unit models for each language can be derived and used for language identification. As a final system, when the performance of the system which uses acoustic likelihood alone is combined with the performance of the unique syllable-like unit models approach, it is shown that the language identification performance of the system improves considerably.

References

Berkling, K.M., Arai, T., Bernard, E., 1994. Analysis of phoneme based features for language identification. In: Proceedings of IEEE Int. Conf. Acoust. Speech Signal Process., April 1994, pp. 289–292.

Fujimura, O., 1975. Syllable as a unit of speech recognition. IEEE Trans. Acoust. Speech Signal Process. 23 (1), 82–87.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: Proceedings of IEEE Int. Conf. Acoust. Speech and Signal Process., pp. 517–520.

Greenberg, S., 1999. Speaking in short hand—a syllable-centric perspective for understanding pronunciation variation. Speech Comm. 29, 159–176.

Hazen, T.J., Zue, V.W., 1994. Recent improvements in an approach to segment-based automatic language identification. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., September 1994, pp. 1883–1886.

Jayaram, A.K.V.S., Ramasubramanian, V., Sreenivas, T.V., 2003. Language identification using parallel sub-word recognition. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., vol. I, pp. 32–35.

Kadambe, S., Hieronymus, J.L., 1995. Language identification with phonological and lexical models. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., May 1995, pp. 3507–3510.

Lamel, L.F., Gauvain, J.L., 1994. Language identification using phone-based acoustic likelihoods. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., April 1994, vol. 1, pp. 293–296.

Li, K.P., 1994. Automatic language identification using syllabic spectral features. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., April 1994, pp. 297–300.

Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Amer. 58 (4), 880–883.

Muthusamy, Y.K., Cole, R.A., Oshika, B.T., 1992. The OGI multilanguage telephone speech corpus. In: Proceedings of Internat. Conf. Spoken Language Process., October 1992, pp. 895–898.

Muthusamy, Y.K., Barnard, E., Cole, R.A., 1994. Reviewing automatic language identification. IEEE Signal Process. Mag., 33–41.

Nagarajan, T., Murthy, H.A., 2004. Language identification using parallel syllable-like unit recognition. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., Montreal, Canada, May 2004, vol. 1, pp. 401–404.

Nagarajan, T., Murthy, H.A., Hegde, R.M., 2003. Segmentation of speech into syllable-like units. In: Proceedings of EUROSPEECH, Geneva, Switzerland, September 2003, pp. 2893–2896.

Nakagawa, S., Hashimoto, Y., 1988. A method for continuous speech segmentation using hmm. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 960–962.

Navratil, J., Zuhlke, W., 1997. Phonetic-context mapping in language identification. In: Proceedings of EUROSPEECH, Greece, September 1997, vol. 1, pp. 71–74.

Noetzel, A., 1991. Robust syllable segmentation of continuous speech using neural networks. In: Electro International, pp. 580–585.

Prasad, V.K., 2003. Segmentation and Recognition of Continuous Speech. Ph.D. Dissertation, Indian Institute of Technology, Department of Computer Science and Engineering, Madras, India.

Prasad, V.K., Nagarajan, T., Murthy, H.A., 2004. Automatic segmentation of continuous speech using minimum phase group delay functions. Speech Comm. 42, 429–446.

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice Hall, New Jersy.

Ramasubramanian, V., Jayaram, A.K.V.S., Sreenivas, T.V., 2003. Language identification using parallel phone recognition. In: WSLP, TIFR, Mumbai, January 2003, pp. 109–116.

Ramasubramanian, V., Jayaram, A.K.V.S., Sreenivas, T.V., 2003. Language identification using parallel sub-word recognition—an ergodic hmm equivalence. In: Proceedings of EUROSPEECH, September 2003, pp. 1357–1360.

Schmidbauer, O., 1987. Syllable-based segment-hypotheses generation in fluently spoken speech using gross articulatory features. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 391–394.

Shastri, L., Chang, S., Greenberg, S., 1999. Syllable detection and segmentation using temporal flow neural networks. In: ICPhS, San Francisco, pp. 1721–1724.

Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A., 2003. Acoustic, phonetic, and dis-

criminative approaches to automatic language identification. In: Proceedings of EUROSPEECH, Geneva, September 2003, pp. 1345–1348.

Wu, S.L., Kingsbury, E.D., Morgan, N., Greenberg, S., 1998. Incorporating information from syllable-length time scales into automatic speech recognition. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., Seatle, WA, May 1998, pp. 721–724.

Yan, Y., Barnard, E., 1995. A approach to automatic langauge identification based on language-dependent phone recognition. In: Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., May 1995, pp. 3511–3514.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. 4 (1), 31–44.