A Text Categorization Approach to Automatic Language Identification

Sheng Gao*, Bin Ma*, Haizhou Li* and Chin-Hui Lee†

*Institute for Infocomm Research, Singapore 119613
†School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

We propose a novel approach to spoken language identification (LID). In this framework, a group of utterances from a particular language is treated as a “spoken document” characterized by a “document vector”. The collection of spoken documents in the training set from the same language forms a specific “language identification category”. An unknown testing utterance to be identified can also be represented as a query vector, such that LID is accomplished just like in the case of associating a text document to a topic. This process is known as text categorization (TC). The key lies in tokenizing speech signals with a set of “key terms” so that their salient patterns and corresponding statistics can be used to discriminate individual spoken languages. To perform LID we can adopt any classifier learning and feature extraction techniques developed in the TC community. When compared with the prevailing parallel PRLM method, the proposed approach achieves a relative error reduction of about 87.5%, and reaches an error rate of 0.2% and 1.54% for 3 and 6 languages, respectively, with queries of about 10 seconds long.

1. Introduction

Automatic language identification (LID) is a task of identifying the language identity corresponding to a given speech signal. It is an important technology in multi-lingual speech recognition, speech translation, and multi-lingual spoken document retrieval.

In the past few decades, many statistical approaches to LID have been developed [1, 2, 3, 4, 5, 6] by exploiting recent advances in language modeling and large vocabulary continuous speech recognition (LVCSR). To distinguish these systems from our proposed framework, they are collectively referred to as a conventional LID paradigm. Estimating and computing acoustic and n-gram probabilities of phonemes [2, 3, 4, 5] is the key in LID design so that a phone recognizer is used to convert utterances into sequences of symbols. Then acoustic and language scores are combined to make a final decision [5]. We can even relax the phone definitions and directly infer some pseudo-subword units using unsupervised learning techniques [7]. Good results have been reported with pseudo-syllable units for LID [6].

To further improve the performance, other information, such as articulatory features [1], lexical knowledge [8] and prosody [6], have also been integrated into LID systems. Zissman [3] experimentally showed that phonetic language models can be more powerful than the MFCC-based GMMs. Therefore fusion of high-level features and good utilization of their statistics are two important research topics for LID. However, it is often difficult to fuse diverse features extracted from multi-resolution analysis, such as long-term language models, inter-dependency among features, semantic features, and high-order statistics of speech units. Some difficulties are: (1) these features provide different degrees of discrimination information; (2) data sparsity is a major problem when estimating high-order statistics; and 3) simple fusion does not work well in conventional LID systems.

In this paper we propose a text categorization (TC) [9] approach to LID. An utterance is treated as a spoken document represented by some detected speech patterns and their statistics. The collection of spoken documents in the training set from a particular language forms the same “language identification topic”. Each spoken document related to a topic is characterized with a high-dimension document vector, with each element of the vector representing a “key term” to describe the specific speech pattern related to the spoken document. An unknown testing utterance can also be represented as a query vector, such that LID is accomplished just like in the case of classifying a text document. The key lies in tokenizing speech signals with a set of such key terms to discriminate individual spoken languages. Any classifier learning and feature extraction techniques developed in the TC community can now be adopted to perform LID. In the following, we design an LID system with patterns defined as phonemes and their co-occurrence, features extracted by a latent semantic analysis, and classifiers trained with a maximal figure-of-merit discriminative learning method. These techniques have been shown effective in designing robust text categorization [10] systems. We will apply them directly to LID and show that the proposed approach achieves a significant error reduction over the prevailing parallel PRLM method [3] on a 3-language LID task, with queries of about 10 seconds long.

2. Automatic language identification

A language can be identified using the information from multiple sources. A simple way is to characterize a spoken language using probability distributions of spectral features, e.g. GMM-based LID [3]. However, it does not work well because low-level features carry little discrimination information. More informative features, such as phonetic n-gram [3], often have greater capacity than MFCCs. In addition to these phonetic n-grams there are other high-level cues and features, from articulatory, prosodic, and semantic analysis, that are useful to discriminate languages. Hereafter, we refer to such high-level features as patterns. With statistical signal processing, it is feasible to automatically detect these patterns of interest from speech signal. These patterns need to be “tokenized” first, and then their models can be trained.

Given a sequence of features $\mathbf{q}_T$, of length $T$, its corresponding language is determined by applying the Bayes
Theorem to the a posteriori probability \( P(l \mid Q') \), followed by a maximum a posterior decision rule formulated as:

\[
\hat{l} = \arg \max_{l \in S} \sum_{\nu} p(Q' \mid W, \lambda^{LM}, \lambda^{AM}) \cdot p(W \mid \lambda^{LM})
\]

(1)

where \( S \) is the set of languages to be identified, \( W \) is a candidate phoneme sequence in \( S \), and \( \lambda^{LM} \) and \( \lambda^{AM} \) are the acoustic and language models for language \( l \). The first term on the RHS of Eq. (1) is the probability of \( Q' \) given its acoustic and language models, the second is a language probability of \( W \), and the last term is the prior probability and can be dropped out because it is often assumed to be equal for all languages in \( S \).

The exact computation in Eq. (1) involves summing over all possible phoneme sequences. In many implementations, it is approximated by the maximum over all sequences in the sum. To find the most likely phoneme sequence, \( \hat{W} \), for each language \( l \), we can use the Viterbi algorithm over a loop grammar without imposing any phonotactic constraints:

\[
\hat{W} = \arg \max_{\hat{W} \in \Omega} p(Q' \mid \hat{W}, \lambda^{AM}) \cdot p(W \mid \lambda^{LM})
\]

(2)

where \( \Omega \) is the set of all possible phoneme sequences for language \( l \). The solution to Eq. (1) can be approximated as:

\[
\hat{l} = \arg \max_{l \in S} \sum_{\nu} p(Q' \mid W, \lambda^{LM}) \cdot p(W \mid \hat{W}, \lambda^{LM})
\]

(3)

The corresponding log likelihood in the first term on the RHS of Eq. (3) is denoted as \( \log p(Q' \mid \hat{W}, \lambda^{AM}) \), while the second term computes the log \( n \)-gram probability of a phone sequence, \( W = \{w_1, w_2, \ldots, w_N\} \):

\[
L(W \mid \lambda^{LM}) = \log p(W \mid \hat{W}, \lambda^{LM}) + \sum_{n=1}^{N} \log p(W_n | w_{n+1}, w_{n+2}, \ldots, \lambda^{LM})
\]

(4)

Now the log likelihood in Eq. (3) can be approximated as:

\[
\log p(Q' \mid \hat{W}, \lambda^{AM}) = \log p(Q' \mid \hat{W}, \lambda^{AM}) - \sum_{\nu} \log p(W \mid \hat{W}, \lambda^{LM})
\]

(5)

where \( \theta \in [0, 1] \) is a language-specific weight [3] to balance acoustic and language scores.

It is clear that the computation cost to solve Eq. (3) depends heavily on the number of language-specific models, \( L \), in \( S \). The prevailing parallel PRLM method [3] can be demanding for a large \( L \). A large training set is also required to train a good performance phone recognizer for each language. This can be difficult for rarely observed languages because speech examples and exact phonetic specifications for them are often expensive to compile. Furthermore each language-specific phone model is usually trained with language-specific speech examples collected in acoustic conditions that can be quite different from those for training data of other languages. This acoustic mismatch induces a potential acoustic score bias in the likelihood computation and decision comparison in Eq. (3), and therefore degrades the LID performance a great deal in adverse conditions. It is also noted that the acoustic scores dominate the decision process in Eq. (3) in the conventional LID paradigm.

We can simplify the above formulation using a single set of phone models. This universal acoustic model can be obtained by pooling all available acoustic training data for all languages and labeling all utterances with a common set of phoneme units, e.g. the International Phonetic Alphabet (IPA). By doing so the acoustic probabilities are no longer used in score comparison and we need to enhance the overall language probability in Eq. (5) as:

\[
L(W \mid \lambda^{LM}) = \sum_{\nu} \log p(W \mid \hat{W}, \lambda^{LM})
\]

(6)

where \( \lambda = (\log p(W_1), \ldots, \log p(W_n), \ldots, \log p(W_n | W_{n+1}, W_{n+2}, \ldots)) \) is a vector with all \( n \)-gram log probability as its elements, and its dimension is equal to the total number of \( n \)-grams patterns needed, and \( \omega \) is a weight vector of equal dimension, with each component representing the contribution of individual \( n \)-gram probabilities to the overall probability in Eq. (6).

3. A TC-based LID framework

TC is a process of classifying a document into some pre-defined categories. A text document is expressed as a string of words or terms. Syntax and semantics related information is often used to distinguish one document from the others. By representing a document with document-specific statistics of terms and their co-occurences, we can transform a discrete document into a high-dimension vector so that vector space techniques can be used to process documents. To improve the expressiveness and reduce the dimension, feature selection and reduction algorithms, such as singular value decomposition of latent semantic indexing [11], have been developed for feature extraction. Many machine learning algorithms, such as support vector machines (SVM) [12], have been proposed to design text classifiers as well.

By adopting a signal common phone set and constructing corresponding universal phone models to decode speech [13], a speech segment can now be tokenized with a set of “key terms” so that their patterns and their statistics can now be used to discriminate individual “spoken document”. The given collection of spoken documents in the training set from a particular language forms the same “language identification category”. An unknown testing utterance to be identified can be represented as a query vector as well, so that LID is performed just like in the case of classifying a text document [9]. We can now adopt any classifier learning and feature extraction techniques developed in the TC and information retrieval communities to improve LID performance. The proposed framework is conceptually illustrated in Figure 1.

**Figure 1** A unified framework for LID

To apply TC to LID, spoken documents are first decoded into a feature term sequence. If salient patterns can be defined and detected from the speech signal, a lexicon with each pattern as a key term can be established. Each spoken utterance can now be converted into a “spoken document vector”. By grouping training vectors of a particular language into a category, the spoken language can be characterized using these patterns and their statistics, and then classified. Thus four key issues must be addressed: (1) definition of salient features and patterns; (2) automatic detection of these distinctive patterns; (3) formation of discriminative vectors.
with these patterns and their statistics; and (4) designing of robust, high-performance text classifiers.

3.1. Tokenization and salient pattern detection

Phone, syllable, and word units are often used as to characterize spoken sentences. Speech cues and attributes, such as distinctive features [14], can also be defined. Other units, such as acoustic subword segments [7], have also been proposed in case expert knowledge is not easily available to transcribe spoken utterances, especially for rarely heard languages. Detection of general speech salient patterns is an ongoing research topic [15]. We will focus our attention on phones. A detailed description of defining a universal set of phones in this study will be given in the experimental setup in Section 4. Techniques developed in LVCSR can be used to decode all speech segments into phone sequences [6]. From these symbol samples, statistics of phones and their n-gram co-occurrences are obtained. Therefore, the first and second issues discussed above are partially resolved.

3.2. Feature extraction: latent semantic indexing

We now represent a spoken document or a spoken query by a vector with its dimension equal to the size of the pattern lexicon or the total number of the salient features. For example, in Eq. (7), even for a moderate universal phone set of size $M = 124$, the dimension will reach $M^2 + M^3 = 15,500$ when only $M$ unigrams and $M^2$ bigrams are used in our experiments in Section 4.

To represent a document for effective indexing and retrieval without using any detailed syntactic description, latent semantic indexing (LSI) was developed (e.g. [11]). In LSI, a normalized entropy quantity is computed by taking into account the entire set of training documents. In principle it is interesting to note that patterns occur often in a few documents but not as often in others give high indexing power for these documents. On the other hand, patterns occur very often in all documents do not exhibit any indexing power. This desirable property makes LSI a common tool for information retrieval, language modeling [11], natural language call routing [16], and text categorization [9]. Here we use LSI for the spoken document representation.

Given a lexicon with $M$ terms and $N$ spoken language documents, a term-document matrix, $H$, is established. SVD [3] is then used to decompose $H$ into a multiplication of three matrices:

$$H = USV^T$$  \hspace{1cm} (8)

$U: M \times R$ left matrix with rows $u_i, 1 \leq i \leq M, \ S: R \times R$ diagonal matrix of singular values $s_1 \geq s_2 \geq \ldots \geq s_R > 0, \ V: N \times R$ right matrix with rows $v_j, 1 \leq j \leq N$.

If retaining only the top $Q$ singular values in $S$ and zero out others ($R-Q$), the LSI dimension could be effectively reduced to $Q$ much smaller than $R$. Any document or query represented by a vector $\tilde{d}$ in the original $M$ dimensional space, can be transformed into an $R$ dimensional vector, $\tilde{v}$:

$$\tilde{v} = \tilde{d}^T US^{-1}$$  \hspace{1cm} (9)

These reduced vectors are then used to perform LID and train all language classifiers, to be discussed next.

3.3. MFoM Learning of text classifiers

In the proposed framework it is often found that a large portion of the salient patterns occur only a few times in the training data. Therefore this high-dimension representation is very sparse. It is important to learn a robust and discriminative text classifier from a training set of a limited size. Based on the linear discriminant expressed in Eq. (6), we use linear classifier for each language.

Here we use a multi-class maximal figure-of-merit (MFoM) learning approach to training all the linear classifiers. It has been shown to outperform SVM for text categorization [10] because it fully takes advantage of both positive and negative training instances. Since all classifiers are trained simultaneously, it is quite robust especially for categories with very limited training samples [10]. More importantly, MFoM learning smoothly embeds any preferred performance metric into classifier design. In this study the performance metric to optimize is the empirical LID error over the training set. This is similar to the minimum classification error (MCE) formulation adopted in ASR [17].

4. Experimental results

In all following experiments we tested on a set of six spoken languages, i.e. English with Asian accent, Korean, Mandarin, Cantonese, Shanghainese, and Japanese. A set of 124 phonemes, a combined set of all the phonemes needed to characterize the first three of the six languages, namely English (44 phonemes), Korean (37 phonemes) and Mandarin (43 phonemes), was used to represent the set of common speech unit for the six languages. We also ignore some potential overlap in definition in this universal phone set.

Each phoneme was modeled by a 3-state, continuous density hidden Markov model, with each state modeled by a mixture Gaussian density with 32 mixture components [3]. In addition, multiple fillers were used to absorb non-speech segments at the beginning and end of all utterances. These models were obtained from a corpus of these three languages (Mandarin, English and Korean), each with about 150-200 hours of data collected from landline and mobile phone conversations, and digitized at a sampling rate of 8 kHz. An adaptive voice activity detector was used to extract all speech segments for further processing. The feature vectors consist of 12 MFCCs, normalized energy, and their first and second order time derivatives. These features are then normalized with utterance-based cepstral mean subtraction.

Two additional data sets, namely T10 and E10, serving as the training set for obtaining statistics of speech units for each language, and E10 serving as the test set for evaluating LID performance, were used for the six languages. For simplicity all utterances in T10 and E10, after speech detection, were concatenated and then segmented into spoken documents with a uniform length of about 10 seconds each. Each language in T10 and E10 has about 2,000 and 500 documents, respectively. All utterances were decoded with the universal set of 124 phone models obtained above, and a uniform unigram phone language model. Although there were errors in continuous phone recognition, we simply compute the unigram and bigram statistics of the decoded phone units corresponding to each spoken document. 15,500 distinct unigram and bigram patterns and their statistics were extracted. A linear classifier, based on the linear discriminant function in Eq. (6) for each language, was trained using the MFoM learning algorithm [10] to minimize the empirical language identification error of the training set, E10.
4.1. LSI-Based feature representation

We first study the effect of LSI-based feature extraction using 4 different feature sets, F1, F2, F3, and F4. The feature F1 uses the raw frequency counts of the original document representation with 15,500 features, and F3 is a normalized version of F1 by the local and global weights (see [3] for details). Furthermore F2 is a LSI-based representation with a reduced dimension of R=1,449, and F4 is a normalized version over F3 with R=1,099. When compared with the original feature dimension of 15,500, the dimensions for F2 and F4 are significantly reduced. Table 1 shows the average language identification error rates over six languages for the 4 feature sets. It is clear that the LSI-based feature sets (F2 and F4) gave significant improvement over the original sets (F1 and F3), and the normalized versions (F3 and F4) outperformed the non-normalized versions (F1 and F2).

<table>
<thead>
<tr>
<th>Error (%)</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Six-language comparison of 4 feature sets

4.2. Comparison with conventional LID

The same individual sets of phones for Mandarin, accented English and Korean were used to build parallel PRLM systems [7] for comparison. Three new phone models were trained using about 10,000 10-sec speech segments from each language in T10, a much smaller size than the set used to train the universal phone model discussed above. The three phonetic language models were trained with these sentences and the sentence set used to train the universal phone models. The weights needed to balance the acoustic and language scores in Eq. (5) were carefully tuned. Table 2 lists the average LID evaluation results on E10 over the subset of three languages. The best error rate was only 0.2% for the proposed TC method using the same models used to obtain the results in Table 1. When compared to the error rate of 1.6% obtained with the parallel PRLM method, the best relatively error reduction was about 87.5%. Even for F1, with only using frequency counts of the phone statistics as features, we achieved a 70.6% relative error reduction over that obtained with PRLM. It is interesting to note that our proposed system uses only the information conveyed in the decoded phone string, and no tuning is required to balance the acoustic and language contributions.

<table>
<thead>
<tr>
<th>Error (%)</th>
<th>Parallel PRLM</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>F1</td>
<td>F2</td>
</tr>
<tr>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Three-language comparisons with parallel PRLM

5. Conclusion

In this paper we propose a novel approach to automatic language identification. Spoken utterances are first treated as documents represented by detected patterns and their statistics extracted from the speech signal. Documents from the same language are now grouped into the same topic category. As a result any feature extraction and classifier learning algorithms developed for text categorization can now be directly applied to solving LID. This TC formulation facilitates a straightforward fusion of multi-level knowledge and exploits high-order statistics of these diverse and yet salient features. Different from the conventional LID paradigm, only a universal set of phone models covering the acoustic space for all languages is needed for phone decoding. This makes acoustic modeling and LID computation affordable. It also makes easy constructing LID classifiers for rarely seen languages under sparse training data limitations. When compared with the prevailing parallel PRLM method for 3 languages, the proposed method attains a relatively error reduction of 87.5%, and reaches an error rate of 0.2%. In the case of 6 languages, an error rate of 1.54% is obtained with our method. The comparison with PRLM for all six languages is not yet made. In the meantime we are now conducting tests on the 12-language, 1996 NIST Language Recognition Evaluation task. Preliminary results showed similar improvements. Details will be reported later.

6. References