An Acoustic Segment Modeling Approach to Automatic Language Identification

Bin Ma[*]*, Haizhou Li*[*]*and Chin-Hui Lee*[†]

[*]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613

[†]School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

{mabin, hli}@i2r.a-star.edu.sg chl@ece.gatech.edu

Abstract

We propose a novel acoustic segment modeling approach to automatic language identification (LID). It is assumed that the overall sound characteristics of all spoken languages can be covered by a universal collection of acoustic segment models (ASMs) without imposing any phonetic definitions. These segment models are used to decode spoken utterances into strings of segment units. The statistics of these units and their co-occurrences are used to form ASM-derived feature vectors to discriminate individual spoken languages. We evaluate the proposed approach on the 12-language, 1996 NIST Language Recognition Evaluation (LRE) task. With testing queries of about 30 seconds long, our results show that the proposed ASM framework reduces the LID error rate quite significantly when compared with the prevailing parallel PRLM method. We achieved an accuracy of 86.1% using a set of 128 3-state ASMs, with each state characterized by a mixture Gaussian density with 32 mixture components.

1. Introduction

Automatic language identification (LID) is a process of determining the language identity corresponding to a given set of spoken queries. It is an important technology in many applications, such as spoken language translation, multilingual speech recognition [1], and spoken document retrieval [2]. In the past few decades, many statistical approaches to LID have been developed [3, 4, 5, 6, 7, 8, 9, 10, 11] by exploiting recent advances in acoustic modeling [8, 9] of phone units and language modeling of *n*-grams of these phones [4, 7]. Acoustic phone models are used in language-dependent continuous phone recognition to convert speech utterances into sequences of phone symbols with phone language models. Then these acoustic and language scores are combined into language-specific score for making an LID final decision [11].

Syllable-like units have also been experimented [6]. To further improve the performance, other information, such as articulatory and acoustic features [3, 12], lexical knowledge [1, 13] and prosody [14], have also been integrated into an LID system. Zissman [11] experimentally showed that phonetic language models can sometimes be more powerful than the MFCC-based generalized mixture models (GMMs) [9]. 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 fundamental 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 always work in conventional LID systems.

A fundamental question arises here that if phone definition is really needed to identify spoken languages. When human beings are constantly exposed to a language without giving any linguistic knowledge, they learn to determine the language identity by perceiving some of the speech cues in the specific language. It is also noted that in human perceptual experiments, listeners with multilingual background often perform better than monolingual listeners in identifying unfamiliar languages [15]. These reasons motivate us to explore useful speech cues for LID along the same line of a recently proposed automatic speech attribute transcription (ASAT) paradigm for automatic speech recognition [16].

In this paper we propose an acoustic segment modeling approach to LID. It is assumed that the sound characteristics of all spoken languages can be covered by a universal set of acoustic units with no direct link to phonetic definitions. Their corresponding models, called acoustic segment models (ASMs) [17], can be used to decode spoken utterances into strings of such units. The statistics of the units and their co-occurrences corresponding to utterances in a training set of a particular language can be used to construct feature vectors to build LID classifiers. For spoken queries, ASM-derived feature vectors can be extracted in a similar manner and then used to discriminate individual spoken languages.

Hidden Markov models (HMMs) [18] are often used to model these acoustic units, and this collection of ASMs can be established from bottom up in an unsupervised manner, and has been used to construct an acoustic lexicon for isolated word recognition with high accuracy [17]. In this study we investigate three key issues related to applying ASMs to LID, namely: (1) acoustic coverage in terms of the number of units in the acoustic inventory needed to model all languages; (2) acoustic resolution in terms of the required model detail for each ASM; and (3) the complexity and discriminative power of the ASM-derived feature vectors. We show that the proposed approach achieves an accuracy of 86.1% on the 12-language, 1996 NIST Language Recognition Evaluation task, with spoken queries of about 30 seconds long, using a collection of 128 3-state ASMs, with each state characterized by a mixture Gaussian density with 32 mixture components.

2. Universal Language Characterization

For LID, a tokenizer is needed to convert spoken utterances into sequences of fundamental speech units specified in a sound inventory. Usually language-specific phones are used. However units that are not linked to phonetic definitions can be more universal, and therefore conceptually easier to adopt. Such acoustic units are thus highly desirable for universal language characterization, especially for rarely observed languages. In the following we describe three methods to establish a universal representation of speech units for LID.

2.1. Augmented Phoneme Inventory (API)

It has been attempted to derive a universal collection of phonemes to cover all sounds described in an international phonetic inventories. In practice, it is a challenging endeavor because we need a large collection of labeled speech samples for all languages. Note that these sounds overlap considerably across languages. One possible approximation is to use a set of phonemes from several languages to form a superset, named augmented phoneme inventory (API). We anticipate a good inventory to phonetically cover as many targeted languages as possible. This method can be effective when phonemes from all the targeted languages form a closed set. Human perceptual experiments have also shown that listeners' LID performance improved by increasing their exposure to each language [15]. Our API-based tokenization was explored by using a set of all 124 phones from English, Korean and Mandarin, and extrapolating them to the other nine languages in the 12-language LRE task. We will discuss experimental results later.

2.2. Acoustic Segment Model (ASM)

The above phone-based language characterization method suffers from two major shortcomings. First, a combined phoneme set from a limited set of multiple languages cannot easily be extended to cover new and rarely seen languages. Second, a large collection of labeled speech data is needed to train the acoustic and language phone models for each language. To alleviate these difficulties, a data-driven method that does not rely on exact phonetic specifications is preferred. This can be accomplished by constructing consistent acoustic segment models [17] intended to cover the entire sound space of all spoken languages in an unsupervised manner as follows:

Step 1: Segment an utterance into *Q* consecutive segments in a maximum likelihood manner (e.g. [17]), with boundaries, $\{b_0, b_1, \ldots, b_Q\}$, that minimize an overall distortion:

$$D(O,Q) = \sum_{q=1}^{Q} \sum_{t=b_{q-1}+1}^{b_q} d(o_t, \mu_q), \quad (1)$$

where $O = (o_1, o_2, \ldots, o_T)$, with o_t representing the *t*-th observation vector, μ_q is a centroid of the *q*-th segment, and $d(o_t, \mu_q)$ is a distortion between o_t and μ_q. A dynamic programming procedure is often used to obtain the segment boundaries efficiently. For stopping criteria, we used two parameters, the average segment length and frame based distortion score, to control the segmentation process.

Step 2: Apply the above segmentation algorithm to all the utterances in the multiple language speech corpora, and cluster them into *J* classes with a *k*-means algorithm.

Step 3: The speech segments in the same class are regarded as acoustically similar. We then train one continuous density HMM for each class, and establish *J* acoustic segment models to represent the overall acoustic space of all languages.

2.3. Phonetically-bootstrapped ASM (P-ASM)

Note that the above data-driven procedure to obtain ASMs often results in unnecessarily many small segments because no constraints were imposed in segmentation. This is especially severe in the case of segmenting a huge collection of speech utterances given by a large population of speakers from different language backgrounds. The API approach uses phonetically defined units in the sound inventory. It has the advantage of having phonetic constraints in the segmentation process. By using API to bootstrap ASM, called P-ASM, we effectively incorporate some phonetic knowledge about a few languages to guide the ASM training process as follows:

Step 1: Carefully select a few often studied languages, such as English and Japanese, typically with a large amount of labeled speech data, train language-specific phone models, and choose some models to form a set of *M* models for bootstrapping.

Step 2: Use these *M* models to decode all training utterances in the training corpora. Assume the recognized sequences as "true" labels.

Step 3: Force-align and segment all utterances using the available set of labels and HMMs.

Step 4: Group all speech segments corresponding to a specific label into a class. Use these segments to re-train an HMM.

Step 5: Repeat Steps 2-4 several times until convergence.

From our preliminary results we found the above P-ASM training process more stable and it performed better than the unsupervised ASM procedure described in Section 2.2.

3. An ASM-Based LID Framework

Given a sequence of feature vectors, *O*, of length *T*, most LID frameworks identify the corresponding language by applying the Bayes theorem to the *a posteriori* probability, *P(O| l)*, followed by a *maximum a posteriori* decision rule as:

$$\hat{l} = \arg\max_{l \in A} \sum_{W} P\left(O \mid W, \lambda_l^{AM}, \lambda_l^{LM}\right) P\left(W \mid \lambda_l^{LM}\right) P(l), \quad (2)$$

where *A* is the set of all languages, *W* is a candidate phone sequence in *A*, and λ_l^{AM} and λ_l^{LM} are the acoustic and language models for language l. The first term on the right hand side of Eq. (2) is the probability of *O* 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 language in *A*.

The exact computation in Eq. (2) involves summing over all possible phone sequences. In many implementations, it is approximated by the maximum over all sequences in the sum by finding the most likely phone sequence, $\hat{W}_l$, for each language *l*, using the Viterbi algorithm:

$$\hat{W}_l = \arg\max_{W \in B_l} P\left(W \mid l, O, \lambda_l^{AM}, \lambda_l^{LM}\right) \propto \arg\max_{W \in B_l} P\left(O \mid W, \lambda_l^{AM}\right), \quad (3)$$

where B_l is the set of all possible phone sequences for language *l*. Now solution to Eq. (2) can be approximated as:

$$\hat{l} \approx \arg\max_{l \in A} [\log P\left(O \mid \hat{W}_l, \lambda_l^{AM}\right) + \log P\left(\hat{W}_l \mid \lambda_l^{LM}\right)]. \quad (4)$$

Although the above discussion is based on multiple sets of acoustic and language models for tokenization, it is easy to simplify the formulation using a single set of universal phone models, e.g. the ASM units discussed in Section 2. In doing so there are two obvious advantages: (1) the computation cost is significantly reduced because only one phone recognizer is needed to decode utterances; and (2) more importantly, the robustness of LID can be improved by removing a score bias induced by applying different phone models trained in different acoustic conditions. However since only language scores are used to discriminate languages, a more powerful feature vector is needed. We discuss one such representation, called latent semantic indexing (LSI) [19], in the following.

3.1. LSI representation of ASM-derived feature vectors

We now represent a spoken utterance or a spoken query by a vector with its dimension equal to the size of the total number of useful features, including the statistics of the units and their co-occurrences. For example, in the case of API and ASM units, even for a moderate set of size *J=128*, the total dimension will reach a total of *M=J+J*J=16512 features* with *J* unigrams and *J*J* bigrams. It is precisely the usage of such high-dimensional vectors that we expect the discrimination capability to improve even only language features are used.

To represent a text document for effective indexing and retrieval without using any detailed syntactic description, LSI-based document representation was developed [4] to reduce the dimension of the vector that represents a spoken query. 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 units occur often in a few documents but not as much in others give high indexing powers for these documents. On the other hand, units occur very often in all documents do not exhibit any indexing power. This desirable property makes LSI a useful tool to explore language-specific cues.

3.2. LID Classifiers on ASM-derived features

To the feature vectors of high dimension and sparse, SVM (support vector machine) is a classifier of natural choice. Improved LID performance based on SVM using a structural risk minimization principle [20] has been reported [8]. For the high-dimensional API and ASM derived feature vectors, the probability distributions for each language are not known. In this study we simply used a popular SVM package - SVM^{light} (V6.01)[1] with linear kernels to obtain pair-wise binary SVM classifiers of the 12 languages. A spoken query of unknown language goes through all the pair-wise binary SVM classifiers. The language that gains most of the winning votes takes all.

4. Experimental Results

In the following we experimentally analyze the performance of the proposed ASM framework for LID. Only unigrams, bigrams, and trigrams of ASM units were used to characterize all spoken language documents. All LID tests were evaluated on the 1996 NIST Language Recognition Evaluation (LRE) task. The LRE corpus consists of telephone conversation of 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. We use the training sets and development sets of the LDC CallFriend corpus[2] as the training data. Each conversation in the training data is segmented into overlapping sessions of about 30 seconds each, resulting in about 12000 sessions for each language. They were used to train the universal acoustic models for ASM units and language classifiers of the 12 languages. The evaluation set consists of 1492 30-sec sessions, distributed over the 12 languages. We treat each 30-sec session as a spoken document. All results were reported on the 1492 test trials.

We first used three languages, English (44 phones), Korean (37 phones), and Mandarin (43 phones), to pull together 124 phones. In addition, four of the general models for speech detection and confidence measure were added to form an API set of 128 units. The acoustic feature vectors were consisted of 12 MFCCs and normalized energy, plus their first and second order time derivatives. These features are then normalized with utterance-based cepstral mean subtraction. Each unit is modeled by a 3-state, continuous density HMM, with each state characterized by a mixture Gaussian density of 32 mixture components. The API-bootstrapped ASM procedure discussed in Section 2.3 was then used to train 128 ASMs. We only did one iteration of ASM training process in our experiments.

4.1. Comparison with conventional LID frameworks

First we compared the proposed framework with the parallel PRLM (P-PRLM) system [11] with and without score fusion [9], and the API system described in section 2.1. The weights of acoustic and language scores were carefully tuned in P-PRLM. Table 1 listed these four sets of error rates. For the API and ASM systems, we computed the normalized counts of the unigram and bigrams and formed the language-based feature vectors of dimensions 15500 and 16512, respectively. The ASM-based system reduced the errors over the P-PRLM system by about 37%. We note that the Equal Error Rate 5.6% for the P-PRLM and 2.7% for the score fusion were reported in [8], but no language identification results were available for comparison.

Table 1 Comparison of error rates (%) for different models

P-PRLM[3]	22.0
P-PRLM & Score Fusion[3]	17.0
API SVM	19.2
P-ASM SVM	13.9

4.2. ASM acoustic resolution

We next study the effect of the acoustic resolution required to characterize the ASM acoustic space. Using the 128 units described above, we built models with state mixture Gaussian densities of 8 and 16 mixture components as well. The error rates for these two systems (8-mix and 16-mix columns) are listed in Table 2 together with the P-ASM SVM results listed in Table 1 (32-mix column).

Table 2 LID error rates vs. ASM acoustic resolution

	8-mix	16-mix	32-mix
Error Rate (%)	16.8	15.9	13.9

4.3. ASM acoustic and linguistic coverage

It is also interesting to investigate the effect of the acoustic coverage required in terms of the number of ASM units needed to characterize the sound space of all spoken languages. Using the same set of 128 units described above, we clustered them into 64 and 32 ASM units according to acoustic similarity, and built 64 and 32 ASMs with state mixture Gaussian densities of 32 mixture components. The error rates for these two systems (64-ASM and 32-ASM

[1] http://svmlight.joachims.org

[2] http://www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone. The overlap between 1996 NIST evaluation data and *CallFriend* database has been removed from training data as suggested in the 2003 NIST LRE website http://www.nist.gov/speech/tests/index.htm

[3] Results extracted from [9]

columns) are listed in Table 3 together with the P-ASM results listed in Table 1 (128-ASM column). The results in the row labeled "Bigrams" used feature vectors of dimension 1056 (J=32) for "32-ASM" and dimension 4160 (J=64) for "64-ASM". They represent a major dimension reduction from 16512 described above for "128-ASM'. It is not surprising to see that the error rates increased drastically this time from 13.9% to 32.6%, when the ASM coverage is reduced by as much as 75%. This agrees with our intuition that we need at least a reasonable number of ASM units large enough in order to cover the sound variation in all languages. It also showed that these reduced-dimension feature vectors greatly impaired the discrimination power of ASM systems.

To look into this property more closely, we also listed results obtained with only unigrams. Comparing the rows labeled "Unigrams" and "Bigrams", it clearly indicates that small-dimension language feature vectors alone (dimensions 32, 64 and 128 for the three systems, respectively) are not enough to discriminate this set of 12 spoken languages. We also included trigrams in the feature vectors for the "32-ASM" system, with an increased dimension of $M=J+J*J+J*J*J$=33824, almost double the dimension of the feature vectors in the best system so far (M=16512). Nonetheless the error rate only improved slightly from 32.6% for the "Brigrams" system to 27.9% for the "Trigrams" case. At this point we did not have the corresponding "Trigrams" results for the 64-ASM and 128-ASM systems due to the feature vector dimensions of these two systems. Some improvements can be expected.

Table 3 Comparison of acoustic and linguistic coverage

Error Rate (%)	32-ASM	64-ASM	128-ASM
Unigrams	40.1	26.7	22.3
Bigrams	32.6	18.6	13.9
Trigrams	27.9	NA	NA

5. Conclusion

We propose a novel acoustic segment modeling approach to automatic language identification (LID). We use a universal collection of acoustic segment models (ASMs) to cover the overall sound characteristics of all spoken languages without imposing any phonetic definitions. These segment models are used to decode spoken utterances into strings of segment units. The statistics of these units and their co-occurrences are then used to form ASM-derived language-only feature vectors to discriminate individual spoken languages. We evaluate the proposed approach on the 12-language, 1996 NIST Language Recognition Evaluation task. With testing queries of about 30 seconds long, an accuracy of 86.1% is achieved using a set of 128 3-state ASMs, with each state characterized by a mixture Gaussian density with 32 mixture components. We also found that acoustic coverage is more critical to maintain than the acoustic resolution for these ASMs. It can be conjectured that we need at least about 100 units in the set of universal acoustic segment units in order to achieve good LID accuracies using only language-derived feature vectors.

6. Acknowledgements

The authors are grateful to Dr. Alvin F. Martin of the NIST Speech Group for making available the 1996 NIST LRE database.

7. References

1. Ma, B., Guan, C., Li, H. and and Lee, C.-H. "Multilingual Speech Recognition with Language Identification," *Proc. ICSLP*, 2002.
2. Dai, P., Iurgel U. and Rigoll, G.. "A novel feature combination approach for spoken document classification with support vector machines," *Multimedia Information Retrieval Workshop,* 2003.
3. Kirchhoff, K., Parandekar, S. and Bilmes, J. "Mixed Memory Markov Models for Automatic Language Identification", *Proc. ICASSP*, 2002.
4. Li, H. and Ma, B. "A Phonotactic Language Model for Spoken Language Identification," *Proc. ACL*, 2005.
5. Matrouf, D., Adda-Decker, M., Lamel, L. F. and Gauvain, J.-L. "Language Identification Incorporating Lexical Information", *Proc. ICSLP*, 1998.
6. Nagarajan, T. and Murthy, H. A. "Language Identification Using Parallel Syllable-Like Unit Recognition", *Proc. ICASSP*, 2004.
7. Parandekar, S. and Kirchhoff, K. "Multi-Stream Language Identification Using Data-Driven Dependency Selection", *Proc. ICASSP*, 2003.
8. Singer, E. *et al*. "Acoustic, Phonetic and Discriminative Approaches to Automatic Language Recognition*," Proc. EuroSpeech,* 2003.
9. Torres-Carrasquillo, P. A. Reynolds, D. A. and Deller. Jr., R. Jr. "Language Identification Using Gaussian Mixture Model Tokenization*", Proc. ICASSP*, 2002.
10. Yan, Y. and Barnard, E. "An Approach to Automatic Language Identification Based on Language Dependent Phone Recognition" *Proc. ICASSP*, 1995.
11. Zissman, M. A. "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *IEEE Trans. Speech and Audio Proc.*, Vol. 4, No. 1, pp. 31-44, 1996.
12. Sugiyama, M. "Automatic Language Recognition Using Acoustic Features," *Proc. ICASSP*, 1991.
13. Adda-Decker, M, *et al*., "Phonetic Knowledge, Phonotactics and Perceptual Validation for Automatic Language Identification", *Proc. ICPhS*, 2003.
14. Hazen, T. J. and Zue, V. W. "Segment-Based Automatic Language Identification", *Journal of Acoustic Society of America*, 101(4):2323-2331, Apr. 1997.
15. Muthusamy, Y. K., Jain, N. and Cole, R. A. "Perceptual Benchmarks for Automatic Language Identification," *Proc. ICASSP*, 1994.
16. Lee, C.-H., "From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition", *Proc. ICSLP*, 2004.
17. Lee, C.-H., Soong, F. K. and Juang, B.-H. "A Segment Model Based Approach to Speech Recognition", *Proc. ICASSP*, 1998.
18. Rabiner, L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE,* Vol.77, No.2, pp. 257-286, 1989.
19. Bellegarda, J. R. "Exploiting Latent Semantic Information in Statistical Language Modeling", *Proc. IEEE*, Vol. 88, No. 8, pp. 1279-1296, 2000.
20. Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, 1995.