Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification*

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

MIT Lincoln Laboratory
Lexington, MA, USA
{es,ptorres,tgleason,wcampbell,dar}@ll.mit.edu

Abstract

Formal evaluations conducted by NIST in 1996 demonstrated that systems that used parallel banks of tokenizer-dependent language models produced the best language identification performance. Since that time, other approaches to language identification have been developed that match or surpass the performance of phone-based systems. This paper describes and evaluates three techniques that have been applied to the language identification problem: phone recognition, Gaussian mixture modeling, and support vector machine classification. A recognizer that fuses the scores of three systems that employ these techniques produces a 2.7% equal error rate (EER) on the 1996 NIST evaluation set and a 2.8% EER on the NIST 2003 primary condition evaluation set. An approach to dealing with the problem of out-of-set data is also discussed.

1. Introduction

Formal evaluations conducted by the National Institute of Science and Technology (NIST) in 1996 demonstrated that the most successful approach to automatic language identification (LID) uses the phonotactic content of a speech signal to discriminate among a set of languages. Phone-based systems, such as those described in [1] and [2], typically employ *n*-gram language models that capture the phonotactics of the token sequences produced by a set of phone recognizers. An alternative approach to LID relies on Gaussian mixture models (GMMs) to classify languages using the acoustic characteristics of the speech signals. Although the GMM approach has been successfully employed for speaker recognition, its language identification performance has consistently lagged that of phone-based approaches [3]. Recently, investigators at Lincoln Laboratory [4] and QUT [5] have described GMM-based language identification systems whose performance matches or exceeds that of phone-based ones.

Both the phone-based and acoustic language recognizers rely on generative techniques to create language models from estimates of underlying class-conditional distributions. A new approach to language identification that uses discriminatively trained support vector machines (SVMs) has been proposed for speaker recognition applications [6] and has been modified for language identification.

This paper presents a description of the Lincoln Laboratory implementation of the phone, GMM, and SVM systems along with evaluations of their performance on standardized tests. Section 2 describes the corpora used in this study and evaluation guidelines followed for the NIST 1996 and 2003 language recognition evaluations. Section 3 presents descriptions of the phone, GMM, and SVM systems as well as that of a recognizer created by fusing the scores of the three systems. A fusion system designed to add robustness in out-of-set conditions is also proposed. Section 4 compares the performance of the systems on the NIST language recognition evaluations, and Section 5 concludes with a brief discussion and summary of the results.

2. Corpora and Evaluation Methods

Experiments reported in this paper were performed in accordance with the NIST 1996 and 2003 language recognition evaluations (LRE), the goals of which were to quantify performance of language identification systems for conversational telephone speech using uniform evaluation procedures [7]. The task in both evaluations was to recognize the language being spoken in speech utterances of three durations (30s, 10s, and 3s) from a set of 12 target languages[1]. Language model training data consists of twenty 30-minute conversations (40 conversations for English, Mandarin, and Spanish) obtained from the Linguistic Data Consortium CallFriend Train set CDROMs [8].

2.1. 1996 NIST LRE

Development data for the 1996 LRE ("lid96d1") consists of approximately 1200 messages for each evaluation duration, with roughly 160 messages each for English, Mandarin, and Spanish and 80 messages for each of the other nine languages. The 1996 LRE test set ("lid96e1") consists of approximately 1500 messages at each duration: 480 for English, 160 each for Mandarin and Spanish, and 80 for each of the other nine languages. English messages were obtained from both the CallFriend corpus (160) and other English corpora (320). The development and test sets were both provided by NIST.

2.2. 2003 NIST LRE

Development data for the 2003 LRE consists of the lid96d1 and lid96e1 sets used in the 1996 LRE. The official evaluation set contains 1280 messages at each of the three durations, with the message breakdown as follows: 80 messages from the CallFriend corpus for each of the 12 target languages, 80 English messages from the Switchboard-1 corpus, 80 English

---
[1] Target languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese.

messages from the Switchboard-cellular corpus, 80 Japanese messages from the CallHome corpus, and 80 Russian messages from the CallFriend corpus. Prior to submission, LRE participants were unaware of the composition of the non-CallFriend material.

2.3. Evaluation methods

System performance is reported as either the language detection equal error rate (EER) or the value of a decision cost function. The overall EER for multiple language identification experiments is computed from the pooled set of all trial scores. The decision cost function (DCF) is given by

$$C_{Det} = 0.5 * (P_{Miss}(\theta) + P_{FalseAlarm}(\theta))$$

For each message, LRE participants were required to make hard decisions (true or false) for each target language so as to minimize the DCF and systems were then ranked based on their DCF scores.

3. Language Identification Algorithms

3.1. Phone-based LID: PPRLM

Lincoln's phone-based LID system is an updated version of the PPRLM (Parallel Phone Recognition and Language Modeling) system described in [1] that was evaluated in the 1996 NIST LRE. PPRLM uses a bank of recognizers to tokenize an incoming message and a set of tokenizer-dependent interpolated language models to score the resulting phone sequences. The six tokenizers are phone recognizers trained from six OGI-TS (Multilanguage Telephone Speech [9]) languages (English, German, Hindi, Japanese, Mandarin, and Spanish) and use null-grammar, three-state, six-mixture HMMs. The updated system uses a 37-dimensional input feature vector (12 cepstra, 12 delta-cepstra, 12 delta-delta-cepstra, and delta-energy) derived from HTK3.1 MFCC coefficients. Non-speech frames are removed using a speech activity detector, and the remaining feature vectors undergo channel normalization using cepstral mean subtraction.

Phone sequences produced by the tokenizers are used to compute gender independent language models (unigram and bigram distributions) for each of the 12 target languages. Training material for the language models is obtained from the 12-language LDC CallFriend Train set. During testing, a vector of 72 language model scores is produced for each test message. Duration-dependent Gaussian backends are trained from the 72-dimensional PPRLM score vectors to produce 12 language scores. Training data for the backend is obtained from the 1996 NIST LRE material. Each score produced by the backend is converted to a log likelihood ratio by dividing it by the average of the other 11 scores.

The phone-based PPRLM system submitted to the 2003 NIST LRE incorporated two additional improvements. First, phones representing silence and closures were added to the symbol set of each language model. Second, trigram distributions were added to the language models, with language-dependent weights for the trigrams, bigrams, and unigrams selected based on development testing.

3.2. Acoustic LID: Gaussian Mixture Models

The GMM LID system described in this paper is an improved version of the one proposed in [4], where high performance was achieved by combining high-order mixture models with shifted delta cepstra (SDC) feature vectors. SDC feature vectors are created by stacking delta cepstra computed across multiple speech frames. SDC computations are controlled by four parameters (N,d,P,k), as discussed in [4]. The 7,1,3,7 SDC parameter configuration (49-dimensional feature vector) selected for this study was based on extensive experimentation conducted by Kohler [10]. Non-speech frames are removed using a speech activity detector and the remaining feature vectors undergo channel normalization using RASTA.

Order 2048 GMM language models are trained for each of the 12 target languages using conversations from the LDC CallFriend Train set. A single background model is trained from the entire train set and language-dependent models are adapted from the background model using the language-specific portions of the data. This method of training allows fast scoring to be used for recognition and was proposed in [5] for LID. Duration-dependent Gaussian backends are trained from the 12-dimensional GMM score vectors to obtain 12 language scores, and these scores are converted to log likelihood ratios, as was done for the PPRLM system, for final evaluation. Training data for the backend is obtained from the 1996 NIST LRE material.

The GMM system submitted for the 2003 NIST LRE incorporated two additional improvements. First, the system used the feature mapping technique proposed by Reynolds [11] to map each feature vector into a channel-independent feature space. The mappings are trained on the Switchboard corpora, as described in [11]. Feature mapping is intended to provide additional robustness to variations in recording conditions and channels. The second modification was the replacement of the gender-independent GMMs with gender-dependent models. These were created by splitting each CallFriend conversation into its two sides and adapting the 24 language-dependent, gender-dependent models from the language-independent, gender-independent background model. For this system, duration-dependent backends were trained from the 24-dimensional score vectors.

3.3. Discriminative LID: Support Vector Machines

Both the PPRLM and GMM language recognizers rely on techniques that estimate underlying class-conditional distributions. Lincoln's newest LID system uses a support vector machine discriminative classifier originally developed for speaker recognition [6]. The SVM uses a Generalized Linear Discriminant Sequence kernel (GLDS) [6] with an expansion into feature space using a monomial basis. All monomials up to degree 3 are used, resulting in a feature space expansion of dimension 22100. A diagonal approximation to the kernel inner product matrix is used, as discussed in [6]. Training material for the language models is obtained from the individual sides of the 12-language LDC CallFriend Train set. Feature vectors are obtained using the same scheme as described for the GMM LID system. The feature vectors are post-processed using short-time feature normalization where each coefficient of each 300 frame block has its mean normalized to 0 and its variance to 1. The feature vector sequence from each conversation side is divided into 5 equal sections, and each section is used to produce an average feature space expansion. After finding the average feature space expansion vectors for all languages, a standard SVM tool (SVMTorch) is used to produce language models using the GLDS kernel. Language model scores for each test

utterance are obtained by computing the inner product between the language model and the average expansion of the utterance. Duration-dependent Gaussian backends are trained from the 12-dimensional SVM score vectors to obtain 12 language scores, and these scores are converted to log likelihood ratios, as was done for the PPRLM system, for final evaluation. Training data for the backend is obtained from the 1996 NIST LRE material.

3.4. Fusion (FUSE3)

The language model scores of the three LID systems were fused using a duration-dependent Gaussian backend classifier. The input vector to the classifier is of dimension 108 (24 GMM scores, 72 PPRLM scores, and 12 SVM scores). The 12 output language scores of the classifier are converted to log likelihood ratios for final evaluation. Training data for the backend is obtained from the 1996 NIST LRE material.

3.5. Multi-corpus and out-of-set LID (FUSE3-OOS)

An experimental system submitted for the 2003 NIST LRE was specifically designed to add robustness for multi-corpus and out-of-set test conditions. The terms multi-corpus and out-of-set are used to refer to message characteristics that are not represented in the training material. As noted in Section 2, a portion of the 2003 LRE test material contained data obtained from non-CallFriend sources (CallHome Japanese and Switchboard English) and out-of-set languages (CallFriend Russian). The identities of the non-CallFriend corpora and out-of-set languages were unknown to the LRE participants prior to the evaluation.

The approach that was taken to provide robustness to unseen conditions was to train a backend Gaussian classifier using the PPRLM, GMM, and SVM scores for the 30-second 1996 LRE messages and for the 45-second "story" files from the OGI-22 corpus [9]. The component systems to the fusion were not modified. The OGI-22 corpus contains messages from 21 languages, 11 of which are common to the NIST-LRE target languages (there are no French messages) and 10 of which are out-of-set. (Russian is one of the OGI-22 languages.) Thus, 11 Gaussians were trained using both 1996 LRE and OGI-22 data, 1 Gaussian (French) was trained from the 1996 LRE material alone, and a 13th Gaussian was trained using all OGI-22 files from the 10 out-of-set languages. Twelve of the 13 output language scores of the classifier were converted to log likelihood ratios for final evaluation.

4. Results

4.1. Evaluation using the 1996 test set

The LID systems described in Section 3 were developed and tested using the NIST 1996 data sets (lid96d1 and lid96e1) described in Section 2. For all systems, duration-dependent, diagonal covariance Gaussian backend classifiers with LDA normalization were trained using lid96d1 to produce 12-dimensional output score vectors. The language identification systems were run on the lid96e1 test sets and the output scores were converted to likelihood ratios by dividing each language score by the average of the other 11 language scores. Results in Table 1 show the equal error rates (in percent) for the systems described in Section 3 along with

Lincoln's 1996 LID system submission. The 95% confidence intervals for these values are approximately ±0.8 (30s), ±1.1 (10s), and ±1.4 (3s).

4.2. Evaluation using the 2003 test set

Results shown in Table 1 indicate that the latest LID systems provide significant performance improvement over Lincoln's 1996 LRE submission. Before being run on the NIST 2003 test data (lid03e1), the backends were retrained using all the available 1996 material (lid96d1+lid96e1). Table 2 shows the equal error rates (in percent) for the submitted systems for NIST's primary testing condition (CallFriend messages only, no Russian). The 95% confidence intervals for these values are approximately ±1.0 (30s), ±1.4 (10s), and ±1.8 (3s). Performance is generally in line with results obtained for the 1996 evaluation data, although several deviations exist.

Table 1: EER (%) performance of Lincoln language identification systems using the 1996 test set.

SYSTEM	30s	10s	3s
PPRLM 1996	9.6	17.8	26.4
PPRLM	5.6	11.9	24.6
GMM	5.1	8.2	16.4
SVM	4.2	11.7	24.0
FUSE3	2.7	6.9	17.4

Table 2: EER (%) performance of Lincoln language identification systems on the 2003 test set for NIST's primary condition (CallFriend messages, no Russian).

SYSTEM	30s	10s	3s
PPRLM	6.6	14.3	25.5
GMM	4.8	9.8	19.8
SVM	6.1	16.4	28.2
FUSE3	2.8	7.8	20.3

Table 3 gives the decision cost function (DCF) values for the systems at each duration. The actual DCF is the value computed from the hard decisions (true or false) assigned to each language hypothesis, and the minimum DCF is the value obtained by minimizing the DCF using truth. Close correspondence between pairs of values indicates that thresholds derived from one set of data can be applied to another relatively accurately.

Table 3: Actual vs. minimum (*italics*) DCF values (%) for the Lincoln language identification systems.

SYSTEM	30s	10s	3s
PPRLM	6.4/*6.4*	13.9/*13.8*	26.0/*25.2*
GMM	4.7/*4.3*	9.8/*9.6*	19.6/*19.5*
SVM	6.0/*5.9*	16.2/*16.1*	28.2/*27.9*
FUSE3	3.0/*2.8*	8.1/*7.5*	20.0/*19.8*

4.3. Multi-corpus and out-of-set evaluation

This section compares the performance of the FUSE3 and FUSE3-OOS LID systems on the 2003 NIST LRE evaluation data (lid03e1). The analysis will focus on the extent to which the FUSE3-OOS system provides robustness to conditions and languages not seen in training and the degree to which

performance is hurt on in-set data. For the FUSE3-OOS LID system, a backend was trained using the lid96d1 and lid96e1 30s messages along with the OGI-22 45s story files. The backend was used in evaluating the 2003 test messages (lid03e1) at all durations. Table 4 compares the performance of FUSE3 and FUSE3-OOS LID for the NIST primary condition (CallFriend messages only, no Russian). The 95% confidence intervals for these values are approximately ±0.9 (30s), ±1.4 (10s), and ±1.7 (3s). Although the 30s and 10s EERs of FUSE3-OOS are worse than those of FUSE3, the differences are statistically insignificant. Thus, the FUSE3-OOS backend has little or no impact on the in-set data.

Table 4: EER (%) performance of FUSE3 systems on NIST's primary condition (CallFriend messages, no Russian) using the 2003 test set.

SYSTEM	30s	10s	3s
FUSE3	2.8	7.8	20.3
FUSE3-OOS	3.6	8.8	20.1

The next set of experiments investigated the performance of both FUSE3 systems on unseen data using an analysis that was consistent with the NIST LRE method of evaluation. The focus of the evaluation was limited to the non-CallFriend English and Japanese utterances and to the CallFriend Russian utterances. Using the development data, duration-dependent thresholds were determined that minimized the decision cost function for each of the two LID systems. The thresholds were applied to the output scores of each language model and produced a true or false decision for each message-model trial for the 2003 test set. Miss and false alarm rates were then computed for these decisions.

Table 5 compares miss and false alarm rates for the 240 non-CallFriend English and Japanese utterances (160 Switchboard English and 80 CallHome Japanese) at each message duration, and Table 6 compares the false alarm rates for the 80 CallFriend Russian utterances. Statistical significance (95% confidence intervals) is given in parentheses. In general, the results indicate that the OOS system degrades performance for in-set data from unseen corpora but significantly reduces the false alarm rates for out-of-set language utterances.

4.4. CPU usage

Twelve-way language identification on a 500 MHz SUN Sparc Ultra-60 runs at the following multiples of real-time: 14 (PPRLM), 0.8 (GMM), and 0.2 (SVM).

Table 5: Miss and false alarm rates (%) for non-CallFriend English and Japanese utterances using true/false decisions determined from the minimum DCF operating points (see text).

DUR	SYSTEM	P_m (%)	P_{fa} (%)
30s	FUSE3	1.7 (0.5–4.2)	1.7 (1.3–2.3)
	OOS	3.7 (1.7–7.0)	1.6 (1.2–2.2)
10s	FUSE3	5.4 (2.9–9.1)	4.9 (4.1–5.8)
	OOS	16.7 (12.2–22.0)	3.2 (2.6–4.0)
3s	FUSE3	15.0 (10.7–20.1)	10.0 (8.9–11.1)
	OOS	37.5 (31.4–44.0)	6.4 (5.5–7.4)

Table 6: False alarm rates (%) for out-of-set Russian utterances using true/false decisions determined from the minimum DCF operating points (see text).

DUR	SYSTEM	P_{fa} (%)
30s	FUSE3	17.9 (15.5-20.5)
	FUSE3-OOS	6.2 (4.8-8.0)
10s	FUSE3	22.7 (20.1-25.5)
	FUSE3-OOS	8.7 (7.0-10.7)
3s	FUSE3	22.1 (19.5-24.8)
	FUSE3-OOS	11.4 (9.4-13.5)

5. Discussion

The results in this paper demonstrate that significant progress has been made in improving the performance of language identification systems. The paper presents performance results on the 1996 and 2003 NIST test sets for three core LID systems, each of which provides performance superior to that of Lincoln's 1996 system. Merging the scores of the core systems using a backend classifier produces further significant improvements, indicating some degree of independence among the decisions being made by the individual systems. The paper also reports on a system designed to achieve language identification robustness to in-set languages obtained from unseen corpora and to out-of-set languages. The approach taken is to focus all robustness-enhancing efforts in the backend rather than in the design of the core systems. Results of the experiment were mixed, with indications that the experimental system could offer some degree of rejection of out-of-set language data but was also more inclined to reject in-set messages from unseen sources.

6. References

[1] M.A. Zissman, "Predicting, diagnosing, and improving automatic language identification performance." *Proc. Eurospeech 97*, Sept. 1997, vol. 1, pp. 51-54.

[2] Y. Yan and E. Bernard, "An approach to automatic language identification based on language-dependent phone recognition." *Proc. ICASSP '95*, vol. 5, May 1995, pp. 3511-3514.

[3] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech." *IEEE Trans. Speech and Audio Proc.*, SAP-4(1), Jan. 1996, pp. 31-44.

[4] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features." *Proc. ICSLP 2002*, Sept. 2002, pp. 89-92.

[5] E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system." *Proc. ICSLP 2002*, Sept. 2002, pp. 93-96.

[6] W.M. Campbell, "A SVM/HMM system for speaker recognition." *Proc. ICASSP 2003*, April 2003.

[7] http://www.nist.gov/speech/tests/lang/index.htm

[8] http://www.ldc.upenn.edu

[9] http://cslu.cse.ogi.edu/corpora/ corpCurrent.html

[10] M.A. Kohler, Personal communication.

[11] D.A. Reynolds, "Channel robust speaker verification via feature mapping." *Proc. ICASSP 2003*, April 2003.