The MITLL NIST LRE 2009 Language Recognition System*

Pedro A. Torres-Carrasquillo, Elliot Singer, Terry Gleason, Alan McCree, Douglas A. Reynolds, Fred Richardson, and Douglas Sturim

Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, MA 02420

{ptorres,es,tgleason,mccree,dar,frichard,sturim}@ll.mit.edu

ABSTRACT

This paper presents a description of the MIT Lincoln Laboratory language recognition system submitted to the NIST 2009 Language Recognition Evaluation (LRE). This system consists of a fusion of three core recognizers, two based on spectral similarity and one based on tokenization. The 2009 LRE differed from previous ones in that test data included narrowband segments from worldwide Voice of America broadcasts as well as conventional recorded conversational telephone speech. Results are presented for the 23-language closed-set and open-set detection tasks at the 30, 10, and 3 second durations along with a discussion of the language-pair task. On the 30 second 23-language closed set detection task, the system achieved a 1.64 average error rate.

1. INTRODUCTION

The National Institute of Science and Technology (NIST) has conducted formal evaluations of language detection algorithms since 1994. The NIST language recognition evaluation (LRE) in the spring of 2009 is the most recent of these evaluations. In this paper, MIT Lincoln Laboratory's primary submission to the NIST LRE09 is discussed. Although the core components of the system are similar to those in recent submissions, the 2009 submission emphasizes reducing the complexity of the system both at the computational and component level while addressing the new challenges in the 2009 evaluation.

The 2009 LRE represents a significant departure from previous evaluations. First, the number of targets has been increased from 14 in 2007 to 23 in 2009 by both introducing new languages and by eliminating the language/dialect distinction. Tasks thus consist of either 23-class closed-set or open set detection. The 23 classes used for this evaluation include: Amharic, Bosnian, Cantonese, Creole, Croatian, Dari, English-American, English-Indian, Farsi, French, Georgian, Hausa, Hindi, Korean, Mandarin, Pashto, Portuguese, Russian, Spanish, Turkish, Ukrainian, Urdu, and Vietnamese. (Although the targets should properly be referred to as classes, this report will follow NIST usage and employ the term "languages.") An additional task of interest in the 2009 LRE is two-class discrimination of confusable language pairs. The confusable pairs considered for the 2009 evaluation included: Bosnian-Croatian, Cantonese-Mandarin, Creole-French, Dari-Farsi, English (American-Indian), Hindi-Urdu, Portuguese-Spanish, and Russian-Ukrainian.

Second, a new emphasis in 2009 was placed on using "found" data for the evaluation rather than conducting collections of telephone speech. Consequently, the conversational telephone speech (CTS) utterances familiar from earlier evaluations were augmented with narrowband telephone segments gleaned from Voice of America (VOA) broadcasts. The addition of VOA material for the evaluation, while eliminating the need for costly data collection, introduced both speaking style and channel variability not present in previous evaluations.

The organization of this paper is as follows. Section 2 describes the core technologies and the new system score fusion and calibration method. Section 3 describes the development data used for the MITLL submission. Section 4 presents the system performance on the NIST 2009 LRE tasks and Section 5 includes a discussion of results.

2. SYSTEMS

The MITLL system submission for the NIST 2009 LRE detection tasks is a combination of three core recognizers: a discriminatively trained GMM spectral system (GMM-MMI), an SVM GMM supervector spectral system (SVM-GSV), and an SVM language classifier using the lattice output of an English tokenizer (EN-SVM). The main components of these systems are described in the remainder of this section along with details of the backend fusion.

2.1 Spectral-based Systems

The spectral-based core recognizers rely on a common feature extraction process. The processing was designed to reduce variability in the signal unrelated to language classification by using variability reduction techniques such as feature normalization, vocal tract length normalization (VTLN), feature-based Wiener filtering, and nuisance attribute projection (NAP).

2.1.1 Feature extraction

The common framework for the spectral systems begins with 20ms windowing of the waveform at a 10ms frame rate processed through a mel-scale filter bank. The output is processed by a

* This work was sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government

RASTA filter whose output is converted into a sequence of cepstral coefficients. The frame-wise cepstra are concatenated into a 56-dimensional feature vector composed of 7 static coefficients and stacked with the set of shifted delta cepstral (SDC) features produced by applying a 7-1-3-7 SDC scheme [1]. The feature vectors are gated against speech activity marks and normalized to a standard normal distribution. In the final step, a feature domain Wiener filtering variation of latent factor analysis spectral compensation is applied to the features of the GMM-MMI system, and feature-domain nuisance attribute projection (fNAP) is applied to the features of the GSV-SVM system. Both methods aim to reduce undesirable variations introduced by low-dimensional sources.

2.1.2 GMM-MMI

The discriminatively trained Gaussian mixture model recognizer developed for LRE09 builds on the system proposed by the group at BUT [2] and is similar to the one employed by MITLL in previous language evaluations [3]. The system uses 2048 mixture components, segment based (>2s) training from a common initial model from which all the target language models were adapted, and 20 training iterations.

2.1.3 SVM-GSV

The SVM GMM supervector system (SVM-GSV) uses a mean+covariance kernel with a push from SVMs to order 1024 GMMs, as described in [4] and [5]. The system is identical to the one implemented in 2007 [3] except that VTLN was not used, a decision based on development testing.

2.2 Tokenizer-based System (EN-SVM)

The SVM 4-gram system uses a discriminative keyword selection approach [6]. From an initial trigram phone SVM system, 4-grams are generated discriminatively using an alternating filter-wrapper algorithm. In the wrapper step, the most discriminative trigrams are selected according to their support vector weights. Then in the filter step a set of 4-grams is created by appending and prepending each selected trigram with a single phone from the phone set. The resulting 4-gram SVM demonstrates a significant improvement in performance over the initial trigram SVM [6].

To deal with the non-language variability caused by the cross conditions between conversational and broadcast news speech as well as within language sources, we applied nuisance attribute projection (NAP) to the 4-gram kernel. A co-rank of 64 was used for the subspace projection. More details on applying NAP to an SVM token system can be found in [7].

2.3 Calibration and Fusion

The backend used for the 2009 evaluation is an extension of the one used in the 2007 [3] and modified for enhanced calibration and accuracy. The backend processing consists of per-system calibration followed by linear fusion. As in [8], a single backend is used for all durations and both open-set and closed-set conditions, where the effect of duration is modeled by a simple parametric function and the closed/open output scores are generated from the same underlying identification posteriors using Bayes' rule. The calibration stage uses a new approach of a single discriminatively-trained (MMI) Gaussian with shared covariance to replace the separate ML training of a Gaussian backend followed by logistic regression calibration. Fusion coefficients are then discriminatively trained using multiclass logistic regression. For the case of the confusable pairs, the language-pair likelihood ratios are produced by rescaling the closed-set identification posteriors for the specific pairs under consideration.

The backend Gaussians were trained using data from both target and out-of-set languages, whereas the GMM-MMI, SVM-GSV, and EN-SVM language recognizers were trained using only data from the 23 target languages. The fused output scores of the core recognizers constituted MITLL's primary submission to the 2009 LRE.

3. DEVELOPMENT DATA

The development corpus used for LRE09 consisted of two main components:

• Conversational telephone speech (CTS) from previous LRE (1996, 2003, 2005, and 2007) as well as data from CallFriend, CallHome, Mixer, OHSU, and OGI-22 collections.

• Broadcast news (BN) narrowband segments from VOA broadcasts.

The VOA data comprised audited 30s segments provided by NIST as well as additional data harvested from the VOA website. The narrowband segments were automatically extracted from VOA shows using an algorithm similar to that used by BUT [9]. The VOA narrowband segments were further processed to automatically remove English segments in non-English shows. Three partitions, training, development and test, were created for corpus. For development and testing partitions, 3s, 10s, and 30s nested segments were created from longer segments when necessary (e.g., when 3/10/30 segments were not available from a previous LRE). For CTS, we started with the train/dev/test partitions from LRE07 development limited to the LRE09 languages, folded the LRE07-dev data into LRE09-train, made LRE07-test the LRE09-dev, and used LRE07-eval for LRE09-test. For VOA dev and test data, we augmented the NIST audited segments with up to 100 narrowband segments having audio durations between 25-55 seconds. A maximum of three segments per VOA show was allowed. For the VOA train data, segments greater than 10 seconds in duration were selected from all shows not used to derive dev/test segments. For each language up to 400 segments per gender were then selected using an iterative scheme to eliminate duplication of speakers. Out-of-set (OOS) dev and test utterances were obtained from CTS and VOA languages that did not overlap with the 23 LRE09 classes.

4. RESULTS

This section presents the performance of the MITLL system described in Section 2 on the NIST LRE09 closed-set, open-set, and confusable pair recognition tasks. NIST officially evaluates system submissions using a decision cost function C_{avg} computed from hard decision errors and a fixed set of costs and priors, as specified in [10]. Results in this section are presented using both C_{avg} and DET plots.

4.1 Closed-set, Open-set, & Language-pair Performance

DET plots for MITLL's submitted system (fused 3-recognizer combination) for the NIST 23-language closed-set and open-set tasks for the 30s, 10s, and 3s LRE09 test segment durations are shown in Figure 1. While it is evident that the open-set task is more challenging than the closed-set, system performance is not

seriously affected. With the release of the 2009 LRE results it became known that 11 of the 17 out-of-set languages in the NIST evaluation test segments overlapped with the 30 out-of-set training data languages. When the backend was retrained without data from the 11 overlapping out-of-set languages, open-set performance at all durations barely changed, indicating that open-set system performance is robust to unseen languages.

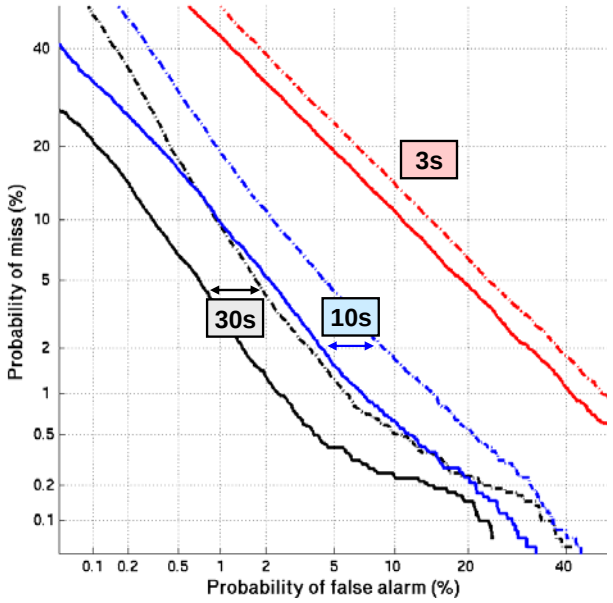

FIGURE 1: ***DET plots for MITLL LRE09 closed-set (solid) and open-set (dashed) system for 30s, 10s, and 3s test segments.***

23-Language CLOSED	SYSTEM	30s	10s	3s
$100*C_{avg}$	SVM-GSV	2.30	5.16	15.61
	EN-SVM	2.34	5.85	17.80
	GMM-MMI	2.45	4.50	13.28
	SVM-GSV + EN-SVM	1.77	3.63	12.41
	SVM-GSV + GMM-MMI	2.00	3.92	11.93
	EN-SVM + GMM-MMI	1.80	3.40	11.20
	ALL	1.64	3.14	10.50

TABLE 1: ***Closed-set performance ($100*C_{avg}$) of individual systems and of combinations. Each configuration used a custom backend.***

Breakouts of closed-set performance of individual core recognizers and combinations for the three test segment durations are given in Table 1. A custom backend was trained for each of the seven system configurations. We observe that closed-set performance of the individual systems is approximately equivalent for the 30s duration segments, whereas the GMM-MMI is superior for the shorter durations. Performance achieved by fusing the scores of the SVM-GSV (spectral) and EN-SVM (token) systems is quite close to that of the submitted 3-way combination.

Closed-set results ($100*$ C_{avg}) for the 8 confusable pairs ranged from 26.3 (Hindi-Urdu) to 0.25 (Portuguese-Spanish). The average for all 253 possible pairs in the 23-language set was 0.59, with 18 pairs having error rates above 1.0. These results indicate that a relatively small number of pairs are highly confusable, and we speculate that the acoustic and phonotactic approaches may be inadequate to address their discrimination.

4.2 Performance by Source: VOA and CTS

The substantially different nature of the speech recorded in conversational telephone environments and telephone-bandwidth broadcast segments raises the question of how well a language recognition system trained on one source type will perform on test data from the other. To evaluate this issue, the SVM-GSV system described in Section 2.1.3 was trained separately from either CTS or VOA training subsets and each was then evaluated on test segments from CTS and VOA. To make the comparison meaningful, data was limited to the 8 classes for which segments from both CTS and VOA were available (Cantonese, Farsi, Hindi, Korean, Mandarin, Russian, Urdu, Vietnamese). All systems used a 2007 LRE version backend. Results for the four possible combinations of train and test conditions for the 30s LRE 2009 test segments are shown in Figure 2. Results indicate that best performance is obtained when a recognizer trained on VOA is tested on VOA, and worst performance (by a factor of more than two) occurs when the same recognizer is tested on CTS. The system trained with CTS data appears more robust to mismatching, with a performance difference of less than 10%. (In fact, this difference is entirely due to the difficulty of distinguishing CTS Hindi and Urdu.) We speculate that the performance discrepancy is due to 1) the difference in speaking styles in the VOA and CTS material, with VOA being characterized by a more structured and less conversational style, 2) the signal properties in the VOA data that may serve to characterize broadcasts by their channels rather than their languages, or 3) the possibility that the same speakers occur in both the train and test partitions of the VOA material.

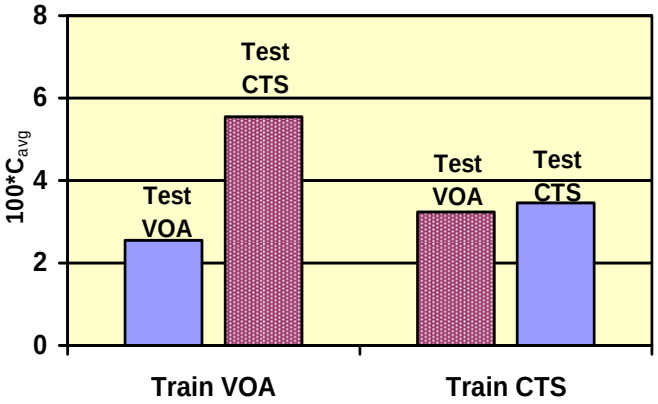

FIGURE 2: ***Performance of SVM-GSV language recognizer on LRE 2009 30s test segments when trained and tested on either VOA or CTS data from 8 languages. Solid: matched case; shaded: mismatched case.***

4.3 Channel compensation

As part of the system development for the 2009 LRE, three compensation techniques (fLFA, fNAP, VTLN) were evaluated in

conjunction with the spectral recognizers (SVM-GSV and GMM-MMI) to determine their effectiveness on the LRE tasks. All three methods were found to be effective for the 2007 LRE (VTLN and fNAP for SVM-GSV, VTLN and fLFA for GMM-MMI) and were consequently employed in the submitted systems [3]. Due to the substantially modified nature of the 2009 LRE tasks, contrastive experiments using the development data were conducted and VTLN was dropped from the 2009 LRE SVM-GSV core recognizer. To evaluate the wisdom of our decisions, contrastive experiments were conducted on the 2009 LRE test segments and results are shown in Figure 3 and Figure 4. It is clear that the greatest gains in performance are obtained using fNAP (SVM-GSV) and fLFA (GMM-MMI), whereas gains using VTLN are smaller and marginal in the case of the SVM-GSV system. We also note that both spectral systems, particularly SVM-GSV, perform quite well in the absence of any additional compensation. We speculate that this may be due to the greatly increased amount of training data available for the 2009 LRE compared to past evaluations.

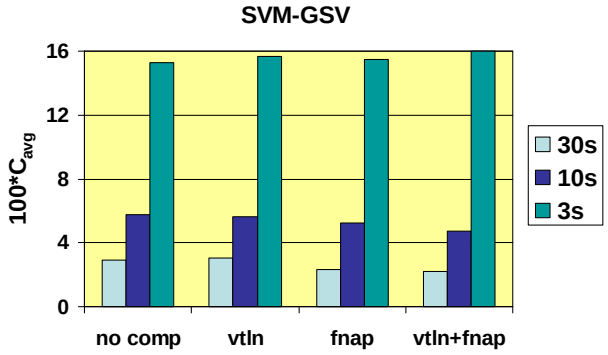

FIGURE 3: ***Effects of channel compensation techniques on the SVM-GSV system for the NIST 2009 LRE closed-set task.***

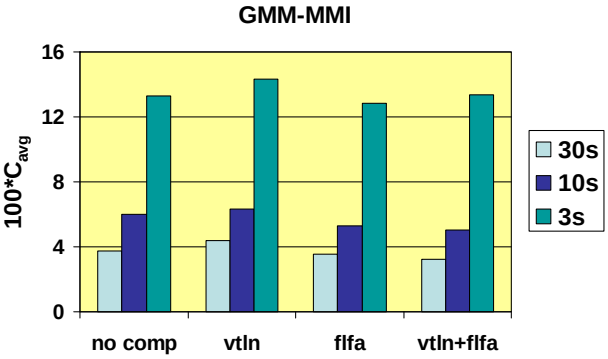

FIGURE 4: ***Effects of channel compensation techniques on the GMM-MMI system for the NIST 2009 LRE closed-set task.***

5. DISCUSSION

Language recognition performance has seen dramatic improvements since NIST began conducting formal evaluations in 1994. This trend has been the product of continued aggressive application of statistical pattern recognition techniques and speech science technology by the speech community to the language recognition problem. Figure 5 shows performance of Lincoln Laboratory systems on the NIST LRE test data beginning with the CallFriend data first employed in 1995-1996, continuing to the more diverse and less structured OHSU and Mixer corpora of 2005-2007, and culminating with the inclusion of broadcast data in 2009. Despite the challenges presented by the more recent collections, technology developers have been able to maintain or improve language recognition performance.

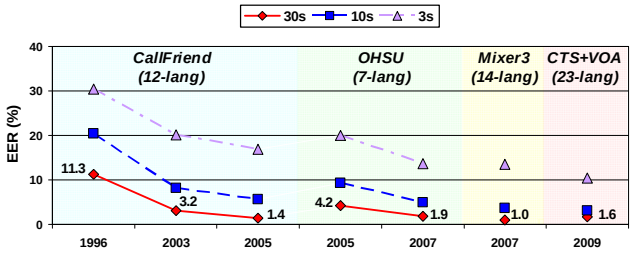

FIGURE 5: ***Performance trends of MITLL language recognition systems on NIST evaluation corpora at three durations. Dates on the horizontal axis indicate the system vintage.***

6. REFERENCES

1. Torres-Carrasquillo, P.A., et al. *Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features*. in *ICSLP*. 2002. Denver, CO.
2. Matejka, P., et al. *BRNO University of Technology System for NIST 2005 Language Recognition Evaluation*. in *IEEE Odyssey: The Speaker and Language Workshop*. 2006. San Juan, PR.
3. Torres-Carrasquillo, P., et al., *The MITLL NIST LRE 2007 Language Recognition System*, in *InterSpeech*. 2008: Brisbane, Australia.
4. Castaldo, F., et al. *Acoustic Language Identification using Fast Discriminative Training*. in *InterSpeech*. 2007. Antwerp, Belgium.
5. Campbell, W.M. *A Covariance Kernel for SVM Language Recognition*. in *ICASSP* 2008. Las Vegas, NV.
6. Richardson, F.S. and W.M. Campbell. *Language Recognition With Discriminative Keyword Selection*. in *ICASSP*. 2008. Las Vegas, NV.
7. Campbell, W. *Compensating for Mismatch in High-Level Speaker Recognition*. in *IEEE Odyssey: The Speaker and Language Workshop*. 2006. San Juan, PR.
8. McCree, A., et al., *Beyond Frame Independence: Parametric Modeling of Time Duration in Speaker and Language Recognition*, in *InterSpeech*. 2008: Brisbane, Australia.
9. Plchot, O., et al. *Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition: Technical Report*, 2009, http://www.nist.gov/speech/tests/lre/2009/radio_broadcasts.pdf.
10. NIST. *LRE-2009 Evaluation Plan*, 2009, http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.