Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Estimating and Exploiting Language Distributions of Unlabeled Data

Alan McCree

MIT Lincoln Laboratory, Lexington, MA 02420
mccree@ll.mit.edu

Abstract

This paper addresses the problem of language distribution estimation from unlabeled data. We present a new algorithm that treats automated classifier identification outputs as likelihoods and iteratively applies Bayes' rule to reclassify the data using successively improving distribution estimates as "priors". Experimental results using the MIT LL submission to the NIST LRE07 evaluation show significant improvements in estimation of non-uniform distributions as compared to a baseline counting approach. In addition, we show how to incorporate these estimated distributions into the classification task. Further experiments on the LRE07 corpus show large gains for both the detection/verification and identification tasks when only a small set of languages are actually present in the test set.

1. Introduction

In many situations, it may be useful to characterize the language distribution of an unlabeled set of speech data using only automated classifier output scores [1]. For example, we may wish to know what percentage of call center customers are speaking Spanish, or which of many data streams is most likely to contain a particular language.

In addition, it may be desirable to utilize this language distribution information to improve the performance of the classifier itself. We would expect that the knowledge that one of two highly-confusable languages is not present in the current dataset would enable higher accuracy in a language identification task.

This paper presents an improved algorithm for distribution estimation using an iterative approach, and experimental results confirming the usefulness of the algorithm for speech data from the 2007 NIST Language Recognition Evaluation (LRE-07) [2]. In addition, we present a straightforward way to incorporate this estimated language distribution into the classifier output using Bayes' rule, and show that this can result in significant improvement in classifier performance in situations where the language distribution is highly non-uniform, in particular when some languages are not present at all.

The paper is organized as follows. First, Section 2 presents a baseline distribution estimation algorithm, our new iterative approach, and experimental performance evaluations. Section 3 then presents our approach for improving language identification and verification performance by utilizing the estimated language distributions, as well as experimental results. Finally, Section 4 presents concluding remarks.

2. Distribution Estimation

2.1. Baseline Approach

The most straightforward way to estimate the language distribution of unlabeled set of speech files is to run a language classifier over all the files, classify each file based on the highest-scoring candidate, and count the number of occurences of each language in the set. The probability distribution estimate for each language can then be written as

$$P(L_i) = n(i)/N \tag{1}$$

where $n(i)$ is number of times language i was chosen and N is the total number of speech files in the test set. This is a generalization of the maximum likelihood distribution estimator for labeled data to the unlabeled case.

This approach is straightforward to implement and understand, and it provides good performance if the language classifier is very accurate. Unfortunately, this distribution can easily be shown to be biased, with the bias becoming more severe for non-uniform distributions and errorful classifiers [1] . An intuitive explanation for this bias is that even if a language is not present at all in the test set, these estimates will not fall below the false alarm rate of the classifier. Therefore, the resulting distribution estimate will be too high for rare classes, and is biased towards a flat distribution.

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.

2.2. Iterative Algorithm

We have developed an algorithm for more accurate estimation of the distribution based on the iterative application of Bayes' rule to the calibrated classifier outputs. Before presenting the algorithm, we first review the use of Bayes' rule and calibration in language identification.

2.2.1. *Bayes' Rule*

If the raw language classifiers estimate the likelihoods of each class for each speech file x, we can use Bayes' rule to estimate the class posteriors for a given prior distribution:

$$P(L_i|x) = \frac{p(x|L_i)P(L_i)}{\sum_j p(x|L_j)P(L_j)} \tag{2}$$

Classification can then be performed on each utterance by selecting the language with the largest posterior.

If the prior distribution of languages is not known, but instead estimated from a training set T, we can write

$$\begin{aligned} P(L_i|x,T) &= \frac{p(x|L_i,T)P(L_i|T)}{\sum_j p(x|L_j,T)P(L_j|T)} \quad (3) \\ &= \frac{p(x|L_i)P(L_i|T)}{\sum_j p(x|L_j)P(L_j|T)} \quad (4) \end{aligned}$$

where we have used the assumption that the current file is not in the training set.

2.2.2. *Calibration*

It is well known that even for a statistical language recognizer based on Gaussian Mixture Models (GMMs), the raw classifier scores do not correspond to the true likelihoods needed for Bayes' rule [3]. Fortunately, we can overcome this problem by using a *back-end* for calibration of the classifier. In particular, identification posteriors generated with equal priors for all classes can be treated as a form of normalized likelihoods. Therefore, a back-end that calibrates equal-prior identification posteriors, such as multiclass logistic regression, can be used for our purpose. Even a detection/verification back-end can be used, where the posterior for each class was generated with a target prior and remaining classes equally distributed. In this case some additional algebra is needed to back-out the target/non-target priors and generate the equal-prior identification posteriors.

2.2.3. *Algorithm*

Eq. 4 suggests an iterative approach to language distribution estimation, where the classification of utterances is updated at each pass using the current distribution estimates as priors in Bayes' rule. This algorithm can be summarized as follows:

- Run a calibrated language identification system over the test set to generate equal-prior identification posteriors (likelihoods) for all classes per file.
- Initialize the language distribution estimate with a uniform distribution.
- For iterations 1 to M:
  - Use Bayes' rule (Eq. 4) to generate identification posterior estimates for each utterance based on likelihoods and the current language distribution.
  - Update the distribution estimate using the counts of each language (Eq. 1).

We typically use $M = 10$ iterations. Notice that the complexity of this algorithm is not high, since it does not require multiple passes of the computationally-expensive raw language identification system. The original scores can be reused in each iteration to generate the updated posterior estimates.

2.2.4. *Example*

To demonstrate the potential performance improvement from this approach, we show a simple example. In this experiment, the true distribution uses only two of the fourteen possible classes. The language identification system is the MIT LL submission [4] for the 2007 NIST Language Recognition Evaluation (LRE). This state-of-the-art system uses a fusion of four individual classifiers, two acoustic and two phonotactic, followed by Gaussian fusion and per-class detection calibration. To make this example challenging to the distribution estimation algorithm, we use the evaluation scores from the relatively errorful short duration (3 second) test condition. This condition produced an equal error rate of 14.4% for the 14-language detection task, with an identification error rate of 38.6%.

As shown in Figure 1, the estimated distribution from the baseline algorithm does not represent the missing classes well. Due to identification errors (particularly false alarms) in the underlying classifier, the counts for the empty classes are too high.

By comparison, the estimated distribution from the iterative algorithm in shown in Figure 2 is much better, zeroing out most of the empty classes.

2.3. Refinements

The proposed iterative estimation algorithm can be viewed as a form of the EM algorithm [5] commonly used in statistical speech processing. In particular, this

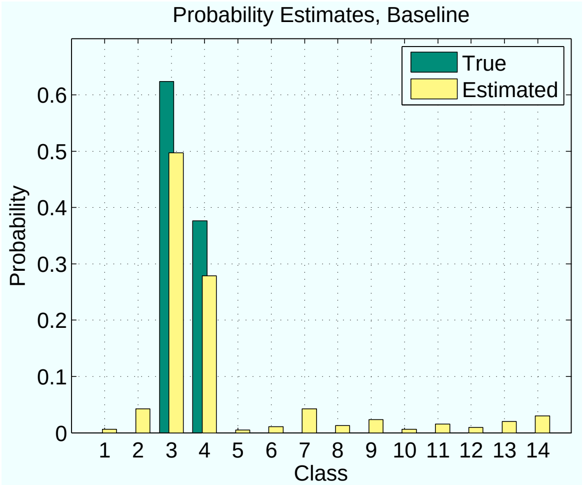

Figure 1: True and estimated language distributions for two-class problem,single pass estimation (baseline).

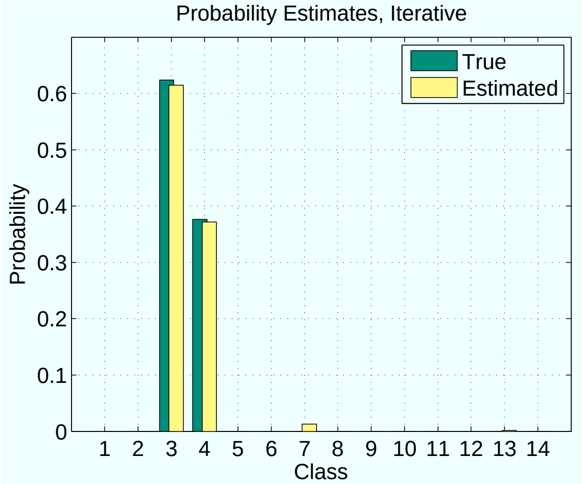

Figure 2: True and estimated language distributions for two-class problem, iterative estimation.

is similar to maximum likelihood mixture weight estimation for Gaussian Mixture Models [6] and to the unsupervised language adaption method proposed in [7]. To see this more clearly, we write the likelihood for each speech file as

$$p(x_n) = \sum_j p(x_n|L_j)P(L_j). \tag{5}$$

In this form, the language distribution plays exactly the same role as the mixture weights in a mixture model, so that the maximum likelihood EM estimation algorithm for GMM weights given in [6] provides a maximum likelihood estimation for the language distribution in our problem. We see that in our iterative language estimation algorithm of Section 2.2.3, the first step in each iteration is the estimation (E), and the second step performs maximization (M).

Based on the GMM EM analogy, we have added two refinements to the distribution estimation algorithm. First, we replace the counts of hard classification decisions of Eq. 1 with the expected counts using the posterior estimates for each class and input file; this leads to a soft count version of the algorithm where

$$n'(i) = \sum_n P(L_i|x_n). \tag{6}$$

In our informal comparisons, this often results in a small performance improvement.

In a second refinement, we address the problem of insufficient training set size by incorporating MAP adaptation of the distribution estimates:

$$P(L_i) = \alpha \frac{n'(i)}{N} + (1-\alpha)P_0(L_i) \tag{7}$$

where

$$\alpha = \frac{N}{N+R}. \tag{8}$$

We use a uniform prior distribution for P_0, and incorporate a small relevance factor R (1 to 10). This has the practical benefit of preventing the algorithm from estimating a zero probability for a class based on only a limited number of observations.

In summary, our refined iterative language estimation algorithm is as presented in Section 2.2.3, but Eq. 1 for updating the distribution estimate is replaced by Eq. 7.

2.4. Performance Evaluation

We have evaluated this distribution estimation algorithm on an extensive sweep of possible distributions. Again, we used the 3 second test scores from the 14-language MIT LL NIST LRE07 submission converted to equal-prior identification posteriors. However, since good calibration is important for the success of the algorithm, we have also evaluated a different back-end combining Gaussian classifiers with multiclass logistic regression as described in [8]. As compared to the MIT LL LRE07 back-end, this newer version provides joint multiclass calibration rather than separate two-class calibration per target, yielding more accurate identification posterior estimates. For these experiments, non-flat true distributions were generated by holding out classes from the testing sets and using the true counts for the remaining classes. The size of the true subset was swept from 2 to 13 classes, and for each subset the results were averaged over 14 random selections of the preserved classes.

2.4.1. Metrics

Two distance measures were used to evaluate the performance of the distribution estimation. The first measure is

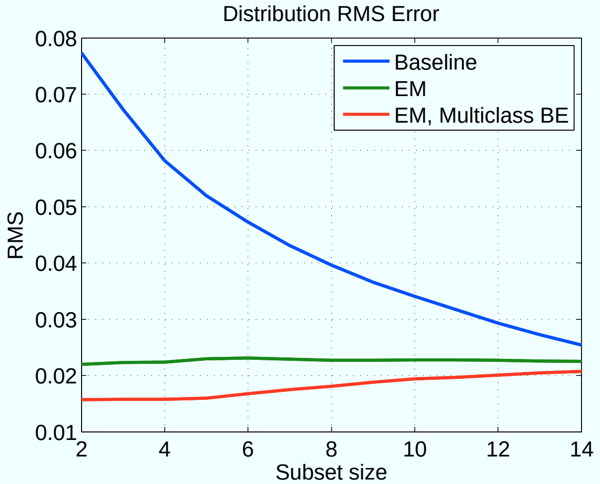

Figure 3: Average RMS error of distribution estimation.

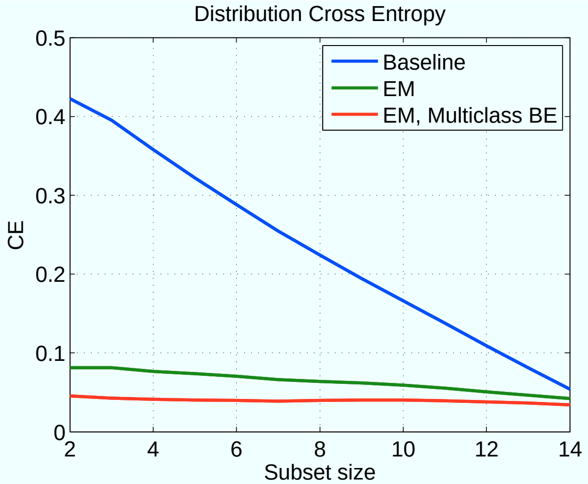

Figure 4: Average cross entropy of distribution estimation.

the RMS error between the true and estimated language distributions, given by

$$RMS = \sqrt{\frac{\sum_i^L (P_i - Q_i)^2}{L}} \tag{9}$$

where P_i represents the true probability of language i and Q_i is the corresponding probability estimate.

The second measure is the cross entropy between the true and estimated distributions, also known as the KL divergence:

$$CE = \sum_i^L P_i \log \frac{P_i}{Q_i}. \tag{10}$$

Note that this measure will highly penalize estimated distributions with incorrect very small values, as the ratio can become extremely large.

2.4.2. *Results*

First, we measured the average error of the distribution estimation algorithms. As shown in Figure 3, the EM algorithm using the submitted system with original back-end provides a significant improvement in RMS error as compared to the single-pass baseline, particularly as the true distribution becomes more uneven with smaller subsets. The bias of the baseline counting approach towards a flat distribution results in increasing estimation error as more classes are not present, while the iterative approach gives comparable error regardless of the subset size in the true distribution. Also, using the multiclass back-end in the EM algorithm provides further improvement, reflecting the benefit of improved identification calibration accuracy.

The results for the cross entropy measure, given in Figure 4, show the same trends.

3. Classifier Performance Improvement

As mentioned in the Introduction, in many situations estimating the distribution is the final goal. However, it is also possible to use these estimated distributions to improve the performance of the underlying classifier in unknown environments. Implicit in the iterative distribution estimation algorithm is the assumption that the classifier can be improved at each iteration by utilizing the distribution estimate as a prior in Bayes' rule. This reduces classifier confusions in the case where one of the two confusable classes is known not to be present in the current test.

Using the same LRE07 language subsets as in the previous section, we measured the potential performance improvement from using these estimated priors to adjust the verification or identification posteriors from the system using Bayes' rule. For these tests, we used the multiclass back-end since it provides better performance. As shown in Figure 5, exploiting the estimated language distributions results in a significant reduction in Equal Error Rate (EER). As the priors become more non-uniform, unlikely candidates are eliminated from consideration, reducing the potential for false alarms. Even the baseline prior estimation algorithm provides considerable improvement, but the EM approach works better and is quite close to actually using the known priors of the test corpus.

Figure 6 shows that the average Bayes' verification cost with a target prior of 0.5, the NIST-defined C_{avg} [2], shows very similar trends. In this case, the theoretical optimal threshold was applied to the verification posterior to make a hard decision for each trial. Finally, Fig. 7 shows the average identification error rate, where the language for each task is estimated by the maximum posterior probability. While identification is fundamentally a

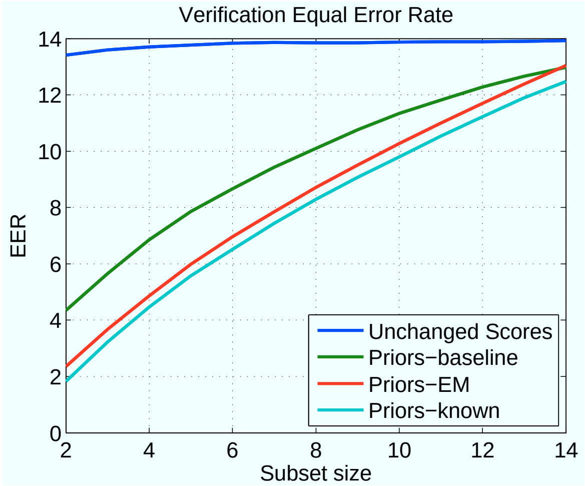

Figure 5: Verification Equal Error Rate.

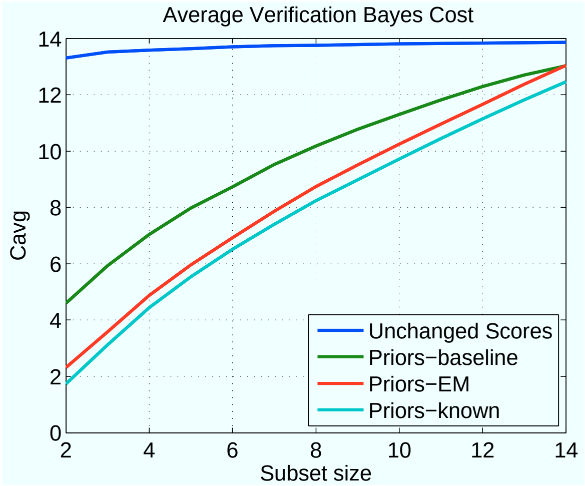

Figure 6: Verification Bayes Cost (NIST priors).

harder task than verification, the trend for this measure is very similar to the previous examples.

4. Conclusion and Future Work

We have presented an improved algorithm to estimate language distributions from unlabeled data. The algorithm treats automated classifier identification outputs as likelihoods, and iteratively applies Bayes' rule to reclassify the data using successively improving distribution estimates as "priors". By sweeping subsets of the NIST LRE07 evaluation corpus, we have shown experimentally that very large improvements can be attained by this algorithm over a baseline counting approach in cases where a significant number of languages are not present.

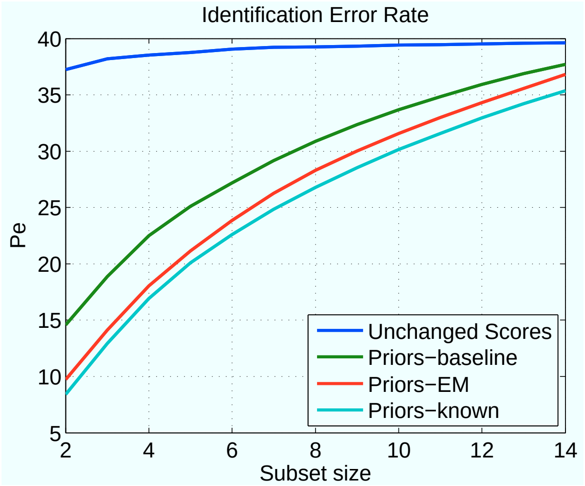

Figure 7: Identification error rate.

In addition, we have shown that incorporation of these estimated distributions into the classification task is straightforward, and can produce big gains in either detection/verification or identification tasks when the raw classifier is errorful and this prior information is strong.

These encouraging results suggest a number of potential areas for future work. First, these experiments have focused on the closed-set task; it would be interesting to know the effect of unknown out of set data on the algorithm performance. Second, these concepts could be extended to include unsupervised adaptation of the backend and/or classifiers to the test set conditions, potentially allowing even greater performance gains. Finally, while these results have been presented for the language distribution estimation task, this algorithm is appropriate for any classification application and for example could be readily applied in speaker identification.

5. Acknowledgements

The author gratefully acknowledges interesting discussions related to this work with Doug Reynolds, John Grothendieck, and Allen Gorin.

6. References

[1] J. Grothendieck and A. Gorin, "Towards link characterization from content: Recovering distributions from classifier output," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, pp. 847–858, May 2008.

[2] "The NIST year 2007 language recognition evaluation plan," http://www.nist.gov/speech/tests/lre/2007/LRE07EvalPlan-v8b.pdf, 2007.

[3] N. Brummer and D. A. van Leeuwen, "On calibration of language recognition scores," in *Proc. Odyssey*, 2006, pp. 1–8.

[4] P. Torres-Carrasquillo, W. Campbell, T. Gleason, A. McCree, D. Reynolds, F. Richardson, W. Shen, E. Singer, and D. Sturim, "The MIT/LL NIST LRE 2007 language recognition system," in *Proc. Interspeech*, 2008, pp. 719–722.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[7] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2003, pp. I224–I227.

[8] A. McCree, F. Richardson, E. Singer, and D. Reynolds, "Beyond frame independence: Parametric modeling of time duration in speaker and language recognition," in *Proc. Interspeech*, 2008, pp. 767–770.