PREDICTING, DIAGNOSING AND IMPROVING AUTOMATIC LANGUAGE IDENTIFICATION PERFORMANCE[*]

Marc A. Zissman

Lincoln Laboratory
Massachusetts Institute of Technology
244 Wood Street
Lexington, MA 02173–9108 USA
Voice: +1 617 981-2547
Fax: +1 617 981-0186
E-mail: `MAZ@SST.LL.MIT.EDU`

ABSTRACT

Language-identification (LID) techniques that use multiple single-language phoneme recognizers followed by n-gram language models have consistently yielded top performance at NIST evaluations. In our study of such systems, we have recently cut our LID error rate by modeling the output of n-gram language models more carefully. Additionally, we are now able to produce meaningful confidence scores along with our LID hypotheses. Finally, we have developed some diagnostic measures that can predict performance of our LID algorithms.

1. INTRODUCTION

We have reported previously that our **P**honeme **R**ecognition followed by **L**anguage **M**odeling performed in **P**arallel (PRLM-P) system provides state-of-the-art language identification (LID) performance on extemporaneous, telephone monologues [5]. In this paper, we wish to report on some recent progress. Section 2 outlines the PRLM-P algorithm including some recent enhancements. We report the performance of this system on conversational, telephone speech in Section 3. In Section 4, some techniques are described that have been used to detect automatically conversations for which our LID hypotheses are likely to be inaccurate. In Section 5, we describe a procedure for analyzing training data to predict whether a given LID problem is likely to be easy or hard even before any test data has been processed. Finally, the paper closes with a review of our conclusions in Section 6.

Applications for LID systems fall into two main categories: preprocessing for machine understanding systems and pre-processing for human listeners. As speech recognition systems proliferate at locations frequented by speakers of many languages (e.g. hotel lobbies, international airports), the LID system would be used as a pre-processor to determine which speech recognition models should be loaded and run. Alternatively, LID might be used to route an incoming telephone call to a human switchboard operator (e.g. emergency or directory assistance) fluent in the corresponding language.

2. LANGUAGE ID ALGORITHM

Described below is the PRLM-P language ID algorithm. Because the basic algorithm has been described in detail elsewhere [5], we only provide a quick summary here. Next, we describe a better way to combine the language likelihood scores, followed by some thoughts on what to do when several sources of training data are available. Finally, we describe an adjunct technique for modeling phone occurrences.

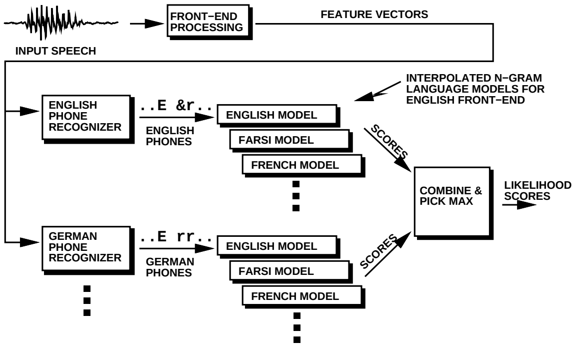


Figure 1. The PRLM-P algorithm.

2.1. Basic Algorithm

Figure 1 shows a block diagram of the PRLM-P system. HMM-based phone recognizers were trained using a phonetically labeled subset of the OGI training speech in each of six languages: English, German, Hindi, Japanese, Mandarin, and Spanish. Each phone recognizer takes as input a stream of mel-weighted cepstra and delta cepstra computed from the incoming digitized speech and produces a stream of phone symbols as output. Interpolated n-gram language models designed to capture the phonotactic statistics of each language are created by passing the training speech for each of the languages to be recognized through each of the six front-end phone recognizers and recording the unigram and bigram counts. We actually create two sets of counts by counting separately those phones that have duration shorter than the mean duration and those having duration longer than the mean duration. We ignore those phones that represent silence and pre-plosive closures. During recognition, the test utterances are passed through each of the phone recognizers, after which the likelihoods of the resulting phone sequences are calculated according to each of the language models. Though we can only build front-end phone recognizers in languages for which we have orthographically or phonetically transcribed speech, we can use the PRLM-P system to perform LID even on languages for which no orthographically or phonetically transcribed speech is available.

2.2. Combining the Scores

The final likelihood scores for each language for each utterance can be calculated any number of ways. The simplest approach is to set the log likelihood that the utterance is spoken in language L equal to the arithmetic sum of the log likelihoods emanating from each of the six language L n-gram models. The underlying assumption of this simple technique for combining the scores is that the various phone recognizers and corresponding language models operate independently from one another. Summing the log likelihoods is equivalent to multiplying linear likelihoods, and multiplying the linear likelihoods is appropriate if events are independent. Although this was our approach through 1995, we have since

[*] THIS WORK WAS SPONSORED BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

adopted the strategy used by Yan [4], whereby we consider the linear likelihoods output by the various language models as elements of a feature vector. If there are N_F front-end phone recognizers and N_L languages to recognize, then there are $N_F \times N_L$ elements in the feature vector. During development testing, we train N_L Gaussian models of the multi-dimensional mean and variance of these likelihood feature vectors. During evaluation of a heretofore unseen utterance, we compute the likelihood given each language and select as our language hypothesis that language whose Gaussian model yields the highest likelihood. We usually use a diagonal, grand covariance matrix, meaning that all models share the same covariance matrix that has non-zero elements only along its diagonal. Presumably we would obtain better performance with language-dependent, full-covariance matrices, but we rarely have enough development test data to estimate so many parameters accurately.

An interesting consequence of combining the scores using this type of Gaussian post-processor is that we can now perform LID for languages for which we have neither front-end phone recognizers nor interpolated language models. As long as we have some development test messages spoken in, say, Arabic, we can create a Gaussian model using a feature vector of likelihoods from n-gram language models that may not include Arabic at all.

2.3. Multiple Sources of Training Data

We are often faced with the problem of having multiple sources of training data. For example, we are now fortunate to have several different multi-language speech corpora available, including OGI_TS (analog collection, monologues) [3], OGI-22 (digital collection, monologues) [2], Linguistic Data Consortium (LDC) CALLFRIEND (digital collection, conversations), and so on. While there is significant language overlap among these corpora, they do not span an identical set of languages. Furthermore, despite our best attempts at normalizing the channels, the fundamental differences in the ways the speech in these corpora was spoken and collected lead our phone recognizers to compute different phone statistics. Therefore, we avoid training a single n-gram language model for language L from the union of language L utterances in these three corpora. Instead, we create one PRLM-P system, including the backend Gaussian classifier described above, for each source of training data. The likelihoods output by the Gaussian classifiers are converted to posterior probabilities, normalized to have zero-mean and unit-variance, and averaged using averaging-weights computed during development test to compute final likelihood scores.

2.4. An Alternative Model for Phone Occurrences

In all of our previous PRLM-P work, we have assumed that phone occurrences can be modeled using a multinomial distribution. Consider a message, M_i, containing a time-ordered sequence of N symbols, s_t, i.e.

$$M_i = \{s_0, s_1, ..., s_t, ..., s_{N-1}\} \qquad (1)$$

The likelihood of this ordered message given a model L_k for language k is:

$$\begin{aligned} Pr_{ord}(M_i|L_k) &= Pr(s_0, s_1, ..., s_t, ..., s_{N-1}|L_k) && (2)\\ &= Pr(s_0|L_k) \times Pr(s_1|s_0, L_k) \times \\ & \quad Pr(s_2|s_0, s_1, L_k) \times \\ & \quad ... \times Pr(s_t|s_0, s_1, ..., s_{t-1}, L_k)... && (3) \end{aligned}$$

We may invoke a unigram approximation, in which we assume that each symbol occurs independently of other symbols.

$$Pr_{ord,unigram}(M_i|L_k) = \prod_{t=0}^{N-1} Pr(s_t|L_k) \qquad (4)$$

Considering this unigram likelihood equation, we note that the ordering of the phones in the sequence is irrelevant, so we may ignore order by defining a count vector, $\vec{n_i}$, whose elements, $n_{i,m}$, are the counts of each type of phone. Assuming we have M types of phones p_m, and letting N_i equal the total number of phone occurrences in message M_i, we can define the multinomial likelihood as:

$$Pr_{mult}(M_i|L_k) = \frac{N_i!}{\prod_{m=0}^{M-1} n_{i,m}!} \times \prod_{m=0}^{M-1} \left(Pr(p_m|L_k)\right)^{n_{i,m}} \qquad (5)$$

These two likelihoods, $Pr_{ord,unigram}$ and Pr_{mult}, are different, but they yield equivalent likelihood ratios, because the leading constant in Pr_{mult} is independent of the model and the data-dependent products are identical. For both models, there is a single parameter to estimate per phone type, namely its mean likelihood of occurrence. This formulation and its extension to bigrams has until now been the foundation of our n-gram language models.

Motivated by the ACQUAINTANCE algorithm developed by Damashek [1], we have recently developed a different method for producing language likelihood scores. Rather than assume a multinomial distribution for the unigrams, we assume no particular source model. We instead create one vector of phone frequencies, $\vec{f_i}$, from each utterance, where the elements of $\vec{f_i}$ are equal to the elements of $\vec{n_i}$ above, except they been each divided by N_i. During training, we can compute a mean frequency vector, $\vec{\mu_k}$ and a covariance matrix, Σ_k, for each language k. During recognition, we can compute a phone frequency vector from a test message and compute the likelihood of that vector given the models for each language using a Gaussian density model, i.e. the probability of observing a vector of phone frequencies, $\vec{f_i}$, from message M_i given a model for language k is:

$$\begin{aligned} Pr_{gauss}(M_i|L_k) = & \quad \tfrac{1}{(2\pi)^{M/2}|\Sigma_k|^{1/2}} \times \\ & \exp\left(-\tfrac{1}{2}(\vec{f_i} - \vec{\mu_k})^T \Sigma_k^{-1} (\vec{f_i} - \vec{\mu_k})\right) \quad (6) \end{aligned}$$

If we assume that the covariance matrix is diagonal, this Gaussian likelihood has two parameters, a mean and a variance, per phone type, as compared to one parameter per phone type available in the multinomial model. Having a second parameter is attractive because we often see more variance in the mean frequency of occurrence of some phones than others. Often the variance is not directly related to the mean, which is a constraint of the multinomial model. The Gaussian model is trained considering each of the training messages individually, rather than considering the data as a whole as in the multinomial case. We call this system "VPF", an abbreviation for "vector of phone frequencies".

An intuitive justification for this VPF model is that the multinomial model for phone occurrences is simply not accurate. Word choice, hence phone occurrence, is certainly dependent on topic and speaker. The putative phone sequences that our recognizers produce are influenced by these factors as well as by noise, channel variability, and speaker dependence that further skew the observed phone frequencies. Thus, the assumption that all phone sequences in some language can be modeled as having been drawn from a single multinomial model is naive. The Gaussian assumption is one way of adding more flexibility to our model.

3. THE 1996 NIST EVALUATION

The National Institute of Standards and Technology (NIST) sponsored its fourth evaluation of LID systems in May 1996. The task was to recognize the language of speech utterances of various durations (3s, 10s, and 30s) from a closed set of 12 possible choices. Three of the 12 languages had two dialects each, but we shall consider those as single languages in this paper. Training data included roughly twenty, 30-minute conversations between friends per language from the LDC CALLFRIEND corpus. The development test set contained four segments for each of the three test durations from a second set of twenty messages per language from CALLFRIEND. The evaluation set contained four segments for each of the three test durations from a third set of twenty messages per language from CALLFRIEND. The evaluation set also

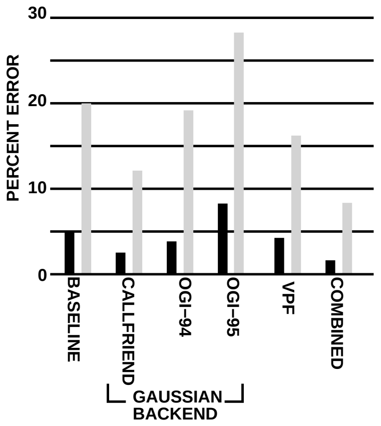


Figure 2. Preliminary performance of the algorithm components on 60s CALLFRIEND segments. Left bar is pairwise classification, right bar is 12-way classification.

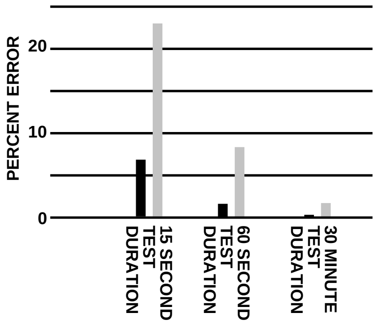


Figure 3. Preliminary performance of the algorithm on CALLFRIEND as a function of duration. Left bar is pairwise classification, right bar is 12-way classification.

contained some additional messages from other American English corpora: KING narrowband NJ, KING narrowband SD, KING wideband, OGI_TS, OGI-22, and SWITCHBOARD. No speaker in either the training or development set appeared in the evaluation set.

Based on some preliminary tests, the Lincoln entry into this evaluation was really five separate LID systems run in parallel. Because all five systems use the same OGI_TS-trained phone recognizers, and because the phone recognition process dominates LID compute time, the use of several different types of postprocessing requires little additional CPU time. The first system used language models trained on CALLFRIEND, the second system used language models *and* a Gaussian backend trained on CALLFRIEND, the third used language models and a Gaussian backend trained on OGI_TS, and the fourth used language models and a Gaussian backend trained on OGI-22. The fifth system was a VPF system trained only on CALLFRIEND. The outputs of these systems were merged by adding the log likelihoods with weights set during development testing. Figure 2 shows cross-validation performance of this combined system and each of its components on 60s test segments of CALLFRIEND speech. We see that the BASELINE system has a higher error rate than one that includes a Gaussian backend classifier. We also see that the combined system out-performs any of the systems operating alone. Figure 3 shows the performance of the combined system on segments of various lengths, from which we see that error rate decreases dramatically as test segment duration increases.

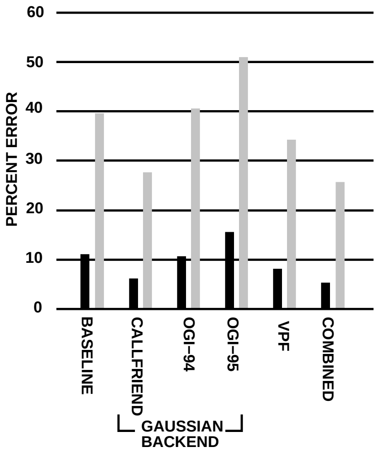


Figure 4. Official 1996 evaluation performance of the algorithm components on 30s CALLFRIEND segments. Left bar is pairwise classification, right bar is 12-way classification.

Performance was measured on the first pass of the algorithm over the 1996 evaluation data, i.e. no tuning on the evaluation set was allowed. Considering only the CALLFRIEND evaluation data, 12-alternative, forced-choice error rate was 25.7% on the 30s cuts, 46.6% on the 10s cuts, and 65.2% on the 3s cuts. Figure 4 shows the error-rate of each of the four component systems individually on the 30s cuts. Again we see the importance of using the Gaussian backend, as the baseline system has a significantly higher error rate than the CALLFRIEND system with the Gaussian backend. We also see that while the combined system yields the lowest error rate, that error rate is just a few absolute percentage points lower than that afforded by the CALLFRIEND system with the Gaussian backend alone.

4. CONFIDENCE SCORES

It is desirable from a practical perspective to know when the LID system is being run over data that is different enough from its training data so as to increase the chance of error. One way to do this is to develop some diagnostic measures that look at a whole group of test messages and measure their similarity to the training data, and this subject will be the topic of Section 5. Another approach is to examine each test message in isolation to measure its similarity to the training data. Of course, that is exactly what the likelihood scores that are output from either the language models or the Gaussian backends measure. But until now, we have always used those likelihood scores to compute likelihood ratios, which purposely mask the overall magnitude of the likelihood scores. Consider an English/German two-language LID problem. If the likelihood scores for both the English and German model are very low, but are different from each other, the likelihood ratio might look reasonable, favoring either English or German, while the raw likelihoods are telling us that neither model fits particularly well. The motivation for computing the likelihood ratio is that we thought we could assume that each test message had to be spoken in either English or German, but when we get two very low raw likelihood scores, we find it may be appropriate to reexamine this assumption.

The 1996 NIST evaluation afforded us the opportunity to test our ability to detect messages that were spoken and/or collected in a different manner from the training data and, hence, might pose a problem for our LID system. While the vast majority of the test messages were drawn from the CALLFRIEND corpus, several hundred English messages from the KING narrowband NJ, KING narrowband SD, KING wideband, OGI_TS, OGI-22, and SWITCHBOARD corpora were also used as test messages. We automatically identified messages that looked “different” by summing the raw likelihood scores output by the Gaussian backends and the VPF system. Our hope was that these summed likelihoods would be low, in general, for messages that were significantly different from the training data. We also suspected that our LID per-

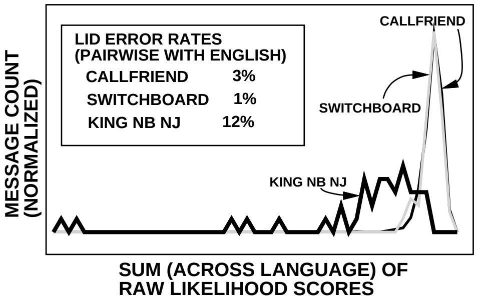


Figure 5. Confidence scores and performance for data from various corpora.

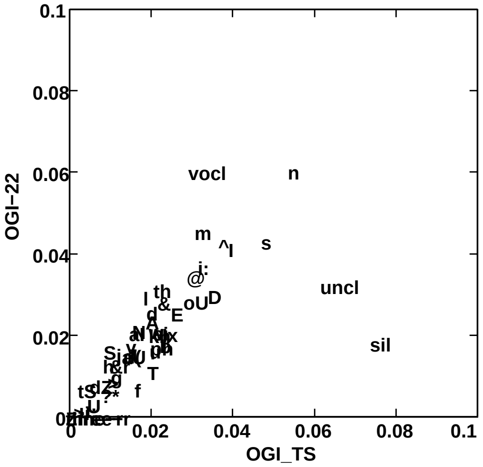


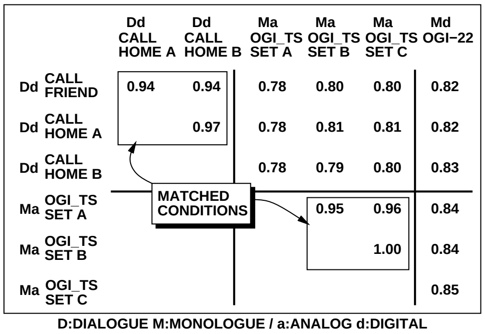
Figure 6. Scatter plot of phone frequency of occurrence between two corpora.

formance on such messages would also be somewhat lower than LID performance on CALLFRIEND messages.

Figure 5 shows the results. The histogram displays message count as a function of summed likelihood for each of three test message corpora. The table shows our LID performance on the same messages (average error rate of pairwise LID of English vs. each of the other 11 languages individually). Results are shown only for the Gaussian backend, but they are similar to that of the VPF system. We see that the KING narrowband NJ messages, which have the worst subjective quality, exhibit both low summed likelihoods and higher LID error rate. Each of the lowest scoring messages is a KING narrowband message, and each was misrecognized by the LID system. SWITCHBOARD messages, which are similar in style and channel to CALLFRIEND messages, yield high summed likelihoods (the SWITCHBOARD and CALLFRIEND histograms are nearly coincident) and an error rate similar to that obtained on CALLFRIEND. We feel hopeful that this summed likelihood measure will continue to be useful in the future as a means for detecting messages that are likely to yield LID errors.

5. DIAGNOSTIC INFORMATION

We are often interested in knowing how well a set of training data matches a set of test data. One approach is to compare the observed frequency of the phones in the training data against that of the test data. As part of the review of the March 1995 NIST evaluation, we did just that, comparing the the phone frequencies in the OGI_TS data (which had been used for training data) to that of the OGI-22 data (which had been used for test data). Yan has also studied these comparisons [4]. In Figure 6, we show a correlation of unigram frequencies, for the English front-end phone recognizer, between the OGI_TS data and the OGI-22 data. These frequencies were computed from messages spoken in all languages of each corpus without knowledge of which messages were spoken in which languages. Figure 7 shows that correlation coefficients between a variety of multi-language corpora. A coefficient of 1.0 would indicate perfect correlation, while a 0.0 would indicate no correlation. Here we can see that the manner of speech (monologue vs. dialogue) and type of collection (analog vs. digital) can each result in mismatched phone frequency statistics. To produce this table, some of the corpora were split into non-overlapping segments so that within corpus correlations could be computed. We include phone frequencies from the LDC CALLHOME corpus, which is another multi-language corpus of conversational telephone speech. As can be seen from the CALLFRIEND, OGI_TS, and OGI-22 results in Figure 4 and the correlations of Figure 7, a mismatch in phone statistics predicts a higher LID error rate. This diagnostic measure can help us predict performance before any LID experiments are run.



	Dd CALL HOME A	Dd CALL HOME B	Ma OGI_TS SET A	Ma OGI_TS SET B	Ma OGI_TS SET C	Md OGI−22
Dd CALL FRIEND	0.94	0.94	0.78	0.80	0.80	0.82
Dd CALL HOME A		0.97	0.78	0.81	0.81	0.82
Dd CALL HOME B			0.78	0.79	0.80	0.83
Ma OGI_TS SET A				0.95	0.96	0.84
Ma OGI_TS SET B					1.00	0.84
Ma OGI_TS SET C						0.85

MATCHED CONDITIONS

D:DIALOGUE M:MONOLOGUE / a:ANALOG d:DIGITAL

Figure 7. Matrix of phone frequency correlations.

6. CONCLUSIONS

We have described a variety of improvements to our LID system. Perhaps the most important was the addition of the Gaussian backend classifier, which cut our error rate by about a factor of two. Other efforts to improve performance, such as running several PRLM-P systems in parallel and using the VPF classifier, provided marginal improvement on the 1996 evaluation data. We have also introduced some message-by-message confidence scores and diagnostic measures for assessing a set of test messages as a group. These two measures can help warn us of potential LID errors.

REFERENCES

[1] M. Damashek. Gauging similarity with n-grams: language-independent categorization of text. *Science*, 267(5199):843–848, February 1995.

[2] T. Lander et al. The OGI 22 language telephone speech corpus. In *Proceedings of Eurospeech 95*, volume 1, pages 817–820, September 1995.

[3] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *ICSLP '92 Proceedings*, volume 2, pages 895–898, October 1992.

[4] Y. Yan and E. Barnard. Experiments for an approach to language identification with conversational telephone speech. In *ICASSP '96 Proceedings*, volume 2, pages 789–792, May 1996.

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