# NIST 2003 Language Recognition Evaluation

*Alvin F. Martin, Mark A. Przybocki*

National Institute of Standards and Technology
Gaithersburg, Maryland, USA
alvin.martin@nist.gov, mark.przybocki@nist.gov

## Abstract

The 2003 NIST Language Recognition Evaluation was very similar to the last such NIST evaluation in 1996. It was intended to establish a new baseline of current performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field. The primary evaluation data consisted of excerpts from conversations in twelve languages from the CallFriend Corpus. These test segments had durations of approximately three, ten, or thirty seconds. Six sites from three continents participated in the evaluation. The best performance results were significantly improved from those of the previous evaluation.

## 1. Introduction

NIST last coordinated an evaluation of language recognition technology in 1996. Late in 2002 NIST announced the plan for a very similar evaluation in early 2003 [1], in order to establish a new baseline of current performance capability for language recognition of conversational telephone speech. Six sites from North America, Europe, and Australia participated. NIST anticipates that the results of this evaluation will establish the groundwork for further research efforts in the field of language recognition.

The evaluation task was to detect the presence of a hypothesized target language, given a segment of conversational speech recorded over telephone lines. Table 1 lists the twelve target languages.

*Table 1:* The Twelve Target Languages

| | | |
|---|---|---|
| **Arabic**<br>(Conversational Egyptian) | **English**<br>(American) | **Farsi** |
| **French**<br>(Canadian French) | **German** | **Hindi** |
| **Japanese** | **Korean** | **Mandarin** |
| **Spanish**<br>(Latin American) | **Tamil** | **Vietnamese** |

## 2. The Evaluation

The performance of a detection system is characterized by its miss and false alarm probabilities. The primary evaluation metric was based upon these. The expected cost of making a detection decision, denoted $C_{Det}$, was defined to be

$$C_{Det} = (C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target}) + (C_{FalseAlarm} \cdot P_{FalseAlarm|Non\text{-}Target} \cdot P_{Non\text{-}Target}) \quad (1)$$

where $C_{Miss}$ and $C_{False\ Alarm}$ represent the relative costs of a miss and a false alarm, respectively. For this evaluation, these were each defined to be 1, and $P_{Target}$, the a priori probability of the target language, was always taken to be 0.5.

The evaluation consisted of a large set of test segments. For each test segment there were twelve trials, corresponding to the twelve target languages.

For each trial, the system provided two outputs. The first was an actual decision ("true" or "false") regarding whether or not the language spoken in the test segment was the target language. The second output was a likelihood score indicating, on an arbitrary scale, how likely it was the test segment language matched the target language.

### 2.1. Data Conditions

The test segments came from one side of a conversation and were represented as standard 8-bit 8 kHz mu-law digital telephone data. Each segment was prepared using an automatic speech activity detection algorithm to identify intervals of speech, which were then concatenated to form the test segment.

The test segments had nominal durations of three seconds (2 s to 4 s), ten seconds (7 s to 13 s), or thirty seconds (25 s to 35 s). They were chosen in sets of three, with each 3 s segment contained within a 10 s segment, which was in turn contained within a 30 s segment. Exactly two such sets of test segments were extracted from each conversation side.

### 2.2. Corpus Support

The primary data source for the evaluation was the multi-language CallFriend Corpus of conversational telephone speech collected several years ago by the Linguistic Data Consortium [2]. This corpus consists of recorded telephone calls made within North America by native speakers of the languages. The languages collected include the 12 specified evaluation languages.

#### *2.2.1. Training Data*

Training data could come from any source. In particular, the 20 complete half hour conversations in each of the 12 target languages from the CallFriend Corpus, which were available for training in the 1996 evaluation, were made available to participating sites in the current evaluation.

#### *2.2.2. Development Data*

Both the development data and the evaluation data from the 1996 evaluation were available as development data for the current evaluation. Each of these sets contained two segments of each duration from each side of 20 CallFriend conversations in each of the 12 target languages.

*2.2.3. Evaluation Data*

The test segments consisted of 80 segments of each duration in each target languages similar to those of the development sets. This data came from conversations collected for the CallFriend Corpus but not heretofore included in the publicly released version of the corpus. In addition, there were four additional sets of 80 segments of each duration selected from other LDC supplied conversational speech sources, namely:

- Russian conversations of CallFriend type
- Japanese conversations from the CallHome Corpus
- English conversations from the Switchboard-1 Corpus
- English conversations from the Switchboard Cellular Corpus

### 2.3. Rules

Participating sites could choose to limit themselves to trials involving only a subset of the twelve target languages. In fact, however, all participants chose to do trials for all twelve languages. They were required to process all 3840 test segments.

The following rules and restrictions applied to all participating sites:

- Each test segment was to be processed separately, independently, and without knowledge of other test segments. Especially, normalization over multiple test segments was not permitted.
- Use of the knowledge of the whole set of target languages was permitted. Thus, normalization over multiple target languages, such as limiting (to say, one) the number of languages for which a "true" decision was made on a given test segment was allowed. Note, however, that there could be, and were, test segments from unknown non-target languages. Use of the knowledge of these languages was not permitted.
- Side knowledge of the sex or other characteristics of the test speaker (except as obtained by automatic means) was not permitted.
- Listening to the evaluation data, or any other experimental interaction with the data, was not permitted before test results were submitted to NIST.

## 3. Detection Performance Results

Figure 1 shows $C_{Det}$ bar charts and Detection Error Tradeoff (DET) curves of performance results for the primary systems of the six participating sites on the thirty second duration CallFriend test segments in the twelve target languages.

The bar charts show both the actual decision $C_{Det}$ values (left bar) and the minimum $C_{Det}$ values (right bar) over all operating points, based on varying the likelihood threshold for decisions. Each bar is divided to show the portions of this cost attributable to false alarms and to missed detections.

The DET plots are ROC-type curves on a normal deviate scale (see [3]). Note that on this scale the plots are approximately linear. The actual decision (♦) and minimum (●) $C_{Det}$ points are shown on each curve. (The site names are omitted, however, as in these evaluations NIST does not publicly identify the individual sites along with their performance results.)

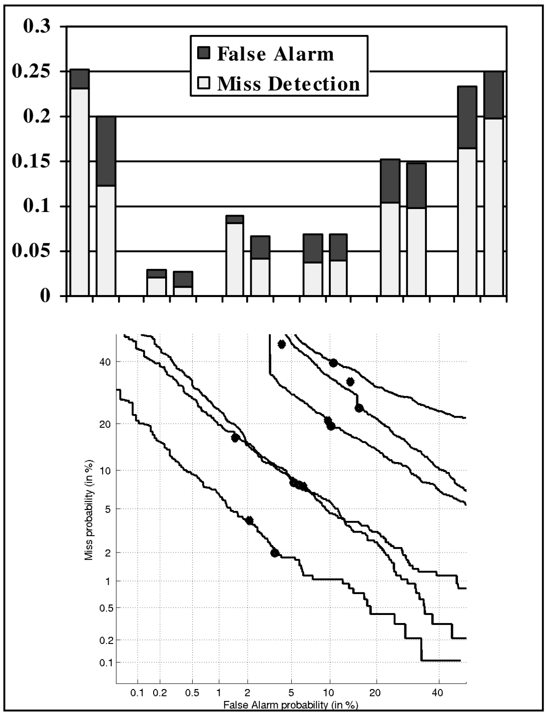

*Figure 1:* $C_{Det}$ bar chart and DET curves for six primary systems on the thirty second duration CallFriend trials in the twelve target languages

### 3.1. Effects of Duration and Sex on Results

Figure 2 shows DET plots of performance results by duration and sex for three of the six primary systems. Not surprisingly, duration is seen to have a major effect on performance for the three durations included in the test set. Further work is needed to determine where the upper limit to this duration effect on performance may lie.

More surprising is the superior performance of most systems on female speech compared to male speech, especially for longer duration test segments. This difference was generally consistent across the different target languages. No such consistency across systems and languages was seen in the 1996 results. The reasons for this apparent performance difference by sex are not apparent, and further investigation seems appropriate.

### 3.2. Comparison with 1996 Results

The previous NIST evaluation of language recognition capabilities in 1996 included CallFriend data in the same twelve target languages similar in type to that used in this evaluation, thus allowing direct comparison of results. Two sites participated in both evaluations. Figure 3 shows DET plots for each duration for these systems in both the 1996 and 2003 evaluations. Clearly, in both cases, there is evidence of considerable performance improvement over seven years.

Participants in the 2003 evaluation will presumably discuss their individual systems in detail elsewhere, but we briefly describe the systems of these two sites here. In the 1996 evaluation both of these sites utilized an approach based on a bank of parallel phone recognizers in multiple languages to tokenize the incoming speech, with language modeling then

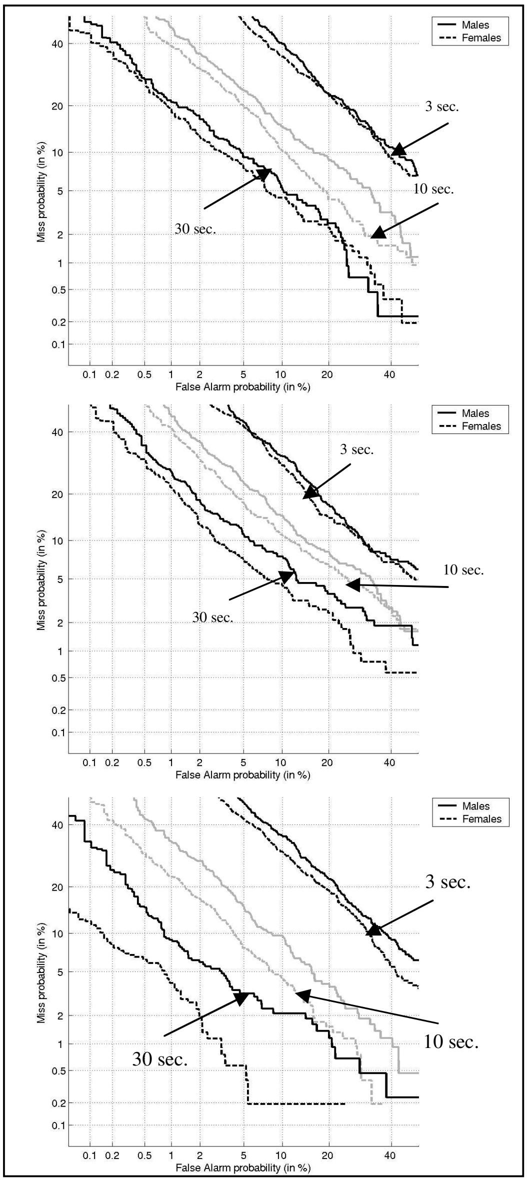

Figure 2: DET plots by duration and sex for systems from three participating sites over all CallFriend trials involving the twelve target languages.

applied to the resulting sets of phone sequences. The languages of the phone sequences need not include all of the target languages. (See, for example, [4].) This was very much the most successful approach at the time, and both sites sought to expand upon it in their 2003 systems.

The site whose system is shown in the upper plot of Figure 3 used an updated version of its previous system, which had six MFCC (Mel-Frequency Cepstral Coefficient) based tokenizers and language models derived from unigram, bigram, and trigram distributions for each of the 12 target languages. This system was fused with two other systems, one using GMM's (Gaussian Mixture Models) and another based on a Support Vector Machine classifier designed for speaker recognition. The combined system ran at about 15 times real-time on a SUN Sparc Ultra-60.

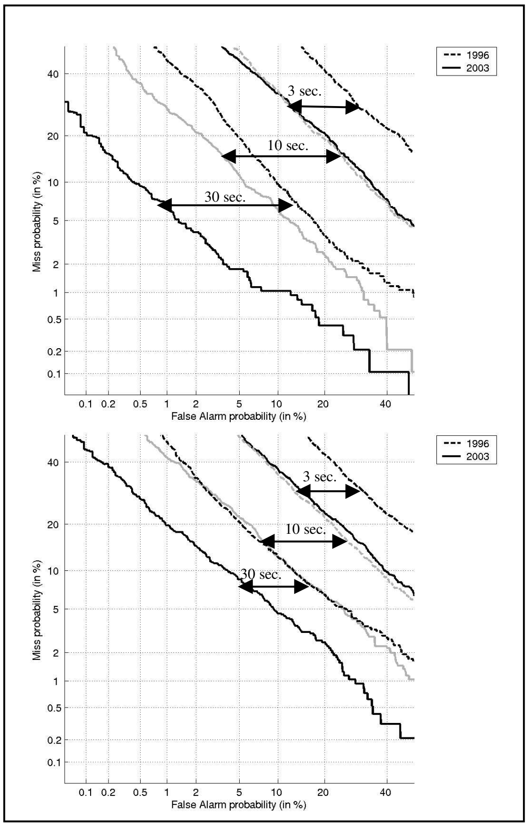

*Figure 3*: Comparison, for two sites participating in both evaluations, of performance over twelve language CallFriend trials in the 1996 and 2003 evaluations for each test segment duration.

The site whose system is shown in the lower plot of Figure 3 also used six language dependent phone recognizers. The system was updated from 1996 to use a trigram rather than a bigram language model and to use MFCC rather than LPC coefficients. Two methods of final classification, one a three layer feedforward neural network and one GMM-based were linearly combined. The total computation time was between one and two times real-time on a Pentium III 933 MHz system.

### 3.3. Language Effects

It is possible to examine the variation in recognition performance by language in several ways. Figure 4 shows performance when the non-target test segments are restricted to one of the thirteen CallFriend language sources (including Russian) while the target trials range over the CallFriend test segments from all twelve announced languages. Results for one system for thirty second duration trials are shown.

With two exceptions, Figure 4 suggests limited performance differences by language. The perhaps expected exception is Russian. With systems not expecting Russian data and not developing models for the language, it is plausible that Russian test segments would prove more difficult to distinguish from the twelve expected languages than segments in these languages. This appears to be the case for the system shown in the figure, and similar results held for other evaluation systems as well.

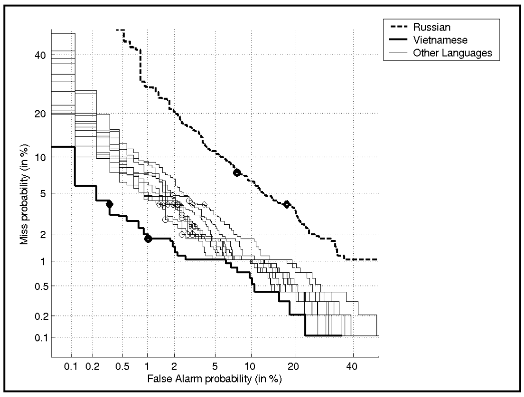

*Figure 4:* DET plots for thirty second duration test segments for one system with non-target trials restricted to CallFriend segments in one of the thirteen languages.

More surprising is that Vietnamese segments appear to be more readily distinguished from other languages than segments in the other target languages. It is not clear why this should be so, but it is a trend that held for all but one of the other evaluation systems. For that system, Mandarin rather than Vietnamese stood out as the easiest of the languages to distinguish. Interestingly, Vietnamese and Mandarin are the two tonal languages among those investigated.

## 4. Identification Results

The evaluation was defined as a detection task, but since the languages of interest were specified in advance and known to the systems, the tests could be combined to do language identification of test segments by selecting the language assigned the maximum likelihood score. Considering the percentages of test segments thus incorrectly identified gives another way to examine performance by language and by data source for the two target languages, English and Japanese, with multiple sources in the test data. Figure 5 shows the percentages of the 80 thirty second test segments of each language and data source thus incorrectly identified by the primary systems of each of the six participating sites. Also shown are the overall error percentages on the 960 CallFriend target language segments of each duration.

Here again Vietnamese is notable as a language with low error rates (as is Mandarin for one particular system). The "easiest" of the data sources, however, was English Switchboard-1, with four of the systems showing zero identification error for it. These conversations are perhaps the cleanest and most clearly spoken of the test data, as speakers did not know each other beforehand and generally stuck to a topic they were assigned beforehand. Note, however, that the English training data provided did not share these properties. The English Switchboard-cellular data also involved unacquainted speakers, but was presumably made somewhat more difficult by the use of cellular handsets. All three English sources were handled fairly well by most of the systems, but the system that scored best for most languages had its greatest difficulty with the English sources other than Switchboard-1

For Japanese, CallHome performance was inferior to that on the similar CallFriend data, perhaps because CallHome calls have one channel originating from outside North America.

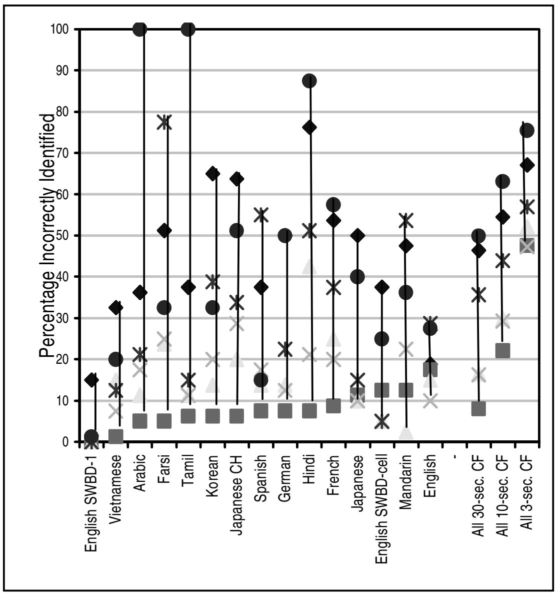

*Figure 5:* Percentages of the thirty second segments incorrectly identified by each primary system for each target language data source (ordered by one system's error percentages). At right are the percentages, by duration, of all target language CallFriend segments incorrectly identified. The sources are CallFriend except where otherwise indicated. The six symbols represent the different primary systems.

## 5. Future Plans

NIST hopes to coordinate similar evaluations in subsequent years. Past experience with other types of evaluation suggests that the techniques used by the best systems in the current evaluation will be incorporated into other systems in future evaluations. Thus a program of regular evaluation can be expected to drive the technology forward.

Several questions raised here may be addressed in future evaluations. These include the apparent better performance with female speech, the range over which performance is sensitive to test segment duration, and the apparent relative ease of distinguishing Vietnamese speech.

NIST evaluations, it should be noted, are open to all research sites that find the task of interest and are willing to discuss their systems at the follow-up evaluation workshop.

## 6. References

[1] "The 2003 NIST Language Recognition Evaluation Plan" http://www.nist.gov/speech/tests/lang/doc/LangRec_EvalPlan.v1.pdf, January, 2003.

[2] Linguistic Data Consortium, Philadelphia, PA, 1996, http://www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone, LDC96S46-LDC96S60.

[3] Martin, A., et al., "The DET curve in assessment of detection task performance", *Proc. EuroSpeeech '97*, Vol. 4 , pp. 1895-1898.

[4] Zissman, M., "Predicting, Diagnosing and Improving Automatic Language Identification Performance," *Proc. Eurospeech '97*, Vol. 1, pp. 51-54.