S10b.4

IMPROVED AUTOMATIC LANGUAGE IDENTIFICATION IN NOISY SPEECH

FRED J. GOODMAN*, ALVIN F. MARTIN*, AND ROBERT E. WOHLFORD**

*GTE GOVERNMENT SYSTEMS
1700 RESEARCH BLVD. ROCKVILLE MD. 20850

**AMERITECH SERVICES
1900 EAST GOLF RD. SCHAUMBURG ILL. 60173

ABSTRACT

This paper describes enhancements to an automatic language identification algorithm previously reported at ICASSP in 1986 [1]. The algorithm, based on LPC-based formant extraction, was greatly improved, reducing the error rate by more than 50 percent. This performance was achieved on a large (>9 hours), very noisy, six-language database, using trials of less than 10 seconds. Experiments that improved performance are described, including tests of various distance metrics, expanded and modified parameter sets, and a new voicing statistic. Final performance results were obtained as a function of time, signal-to-noise ratio, and no-decision rate. A new rejection capability was also developed to address the open-set identification problem.

INTRODUCTION

Although automatic language identification is useful as a front end for other speech recognition devices, it is a problem which has been studied infrequently in the past. Until the work of Foil [1], virtually all research efforts used laboratory quality speech and relatively long speech segments. Foil's work showed the feasibility of recognizing languages with short duration, noisy speech segments. His formant-cluster (spectral peak) approach produced 64 percent recognition with 11 percent no-decisions on a three-language database. This became the starting point for our investigation.

The formant-cluster algorithm used an LPC-12 autocorrelation analysis front end. After LPC coefficients were determined for each 16 millisecond frame (8 kHz sampling rate), an FFT was used to compute the spectrum of the inverse filter. A peak-picking procedure then determined formant frequencies. The first four spectral prominences were defined to be the first four formants [2]. This simplified view avoids unwieldy rule-based formant trackers, yet it retains the concept that such peaks contain vital information, even in high noise.

The parameters used as inputs to the distance classifier were F1/F3, F2/F3, and F4/F3. A Euclidean distance measure determined the closest of the 10 clusters (for each language) to the input vector. That minimum distance was then accumulated. After several seconds a decision was made based on the smallest accumulated distance.

DATABASE QUALITY

The quality of the database was characterized by signal-to-noise ratio (SNR) and by a new "channel grade" measure. The SNR was computed as follows: in every active second, the 95th percentile energy value is divided by the fifth percentile value. The channel grade is defined as the third highest "voicing statistic" (discussed in the next section) per active second. "Active" is defined as being above a voicing statistic threshold. The database quality and size are shown in Table 1.

ENHANCEMENT EXPERIMENTS

The first experiment conducted was designed to determine whether the algorithm was sensitive to channel bandwidth. Therefore, a subband filter was developed to ensure a common channel for all languages. It was implemented as a 100 tap FIR bandpass filter with a 300-2850 Hz passband. After filtering, performance deteriorated about 10 percent, indicating a serious sensitivity. We then eliminated the F4/F3 term and obtained virtually identical results. The subband filter then became a part of the algorithm.

Table 1. Database Summary

Six Languages Database			
	TRAINING SET	RECOGNITION SET	FINAL TEST SET
AVERAGE SNR (dB)	8.99	9.26	9.23
AVERAGE CHANNEL GRADE	86.62	89.16	87.67
TOTAL AUDIO (SECS)	10509	10043	14034
ACTIVE SPEECH (SECS)	3915	4765	5939

CH2673-2/89/0000-0528 $1.00 © 1989 IEEE

The next series of experiments were directed at the classifier. In particular, the Euclidean distance measure was replaced by a weighted Euclidean measure, with the weights being set as the inverse standard deviation of the parameters. This improved performance several percent, and led to experiments with a full Mahalanobis matrix distance measure. A pooled Mahalanobis matrix was derived from 30,000 training vectors (5000 from each of the languages). The pooled matrix proved to be superior to language-specific matrices.

The original parameter set was then enriched by adding terms indicating the log amplitude values at the formant frequencies. The amplitude terms were first computed as ratios A1/A3 and A2/A3. This proved to be inferior to simply using A1, A2, and A3 as separate entities.

This suggested that the division by F3 (done to provide a degree of vocal tract normalization) might actually be hurting performance. Using F1, F2, and F3 as separate terms (expressed on the Bark scale) proved to be superior.

Past results in keyword spotting [3] indicated that 20-40 millisecond time differences on signal processing parameters improved recognition. Time difference terms effectively measure the formant transitions between significant phonetic events for each language. Therefore, 32 millisecond (two frames) time difference terms on the formant frequencies and amplitudes were added. This proved extremely successful. A 64 millisecond time difference proved to be inferior.

The previous algorithm used a voiced/unvoiced decision algorithm which made a significant number of false voicing errors. An improved voicing decision was obtained using a correlation-based voicing statistic to determine active speech. This statistic is successful in rejecting noise, though strong tones must be eliminated prior to processing. Observing the voicing statistic, we realized that it varied significantly within the speech signal, with open vowels scoring highest, voiced fricatives scoring lower, and nasals scoring lower yet. Therefore, the voicing statistic and its time difference were added as parameters. These modifications resulted in a considerable performance improvement.

In training the algorithm, we continued to use the k-means clustering algorithm used in the earlier effort; however, several changes in approach were tested and incorporated. The most important of these changes was to split the training data into "clean" and "noisy" vectors, producing cluster centers for both. During recognition both cluster sets were scored. Finally, after varying the number of clusters for each set from 10 to 50, 30 clusters proved optimum. Thus 60 clusters define each language. The final recognition algorithm is shown in Figure 1.

OTHER EXPERIMENTS

A major alternative to the formant based algorithm was examined to determine the unique contribution of the formant parameters. This algorithm was based on LPC-derived cepstral parameters, which have been successfully used in many speech recognition contexts. The algorithm used a 25 millisecond window with a 12.5 millisecond frame update. The first nine cepstral parameters were used (C0 left out), along with time difference terms from two frames (25 milliseconds) earlier. The voicing statistic discussed earlier was used as well. Tests were run on a 20-parameter algorithm, using varying numbers of cluster centers, and a weighted Euclidean distance measure. The results for the cepstral algorithm were about 10 percent worse than the formant method.

Figure 2 compares performance among several versions of the cepstral and formant algorithms. The formant results shown include a 9 parameter test, (2 formant ratios, 2 amplitude ratios, their time differences, and the voicing statistic), a test after the "clean-noisy" training data split, a 10 parameter test (the voicing statistic time difference was added), a 14 parameter test (eliminating the formant and amplitude ratios), and the final algorithm. The cepstral algorithms shown include using just the basic 9 cepstral parameter test, the 20 parameter approach, and a test using 50 clusters instead of 30.

Experiments were also done to determine whether pitch information (prosody) is useful in performing language identification in such noisy conditions. In limited experiments, accurate pitch tracks were very hard to obtain, even when a dynamic programming tracker was used. The use of syllabic rate as a language discriminant was also briefly investigated. Although the high noise levels made an energy-based approach impossible, use of the voicing statistic for syllable detection appeared to be promising.

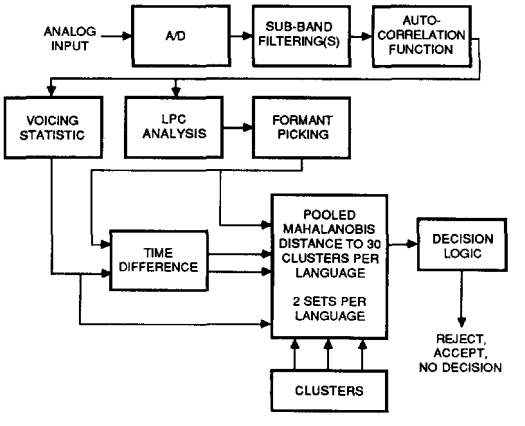

Figure 1. Final Recognition Algorithm

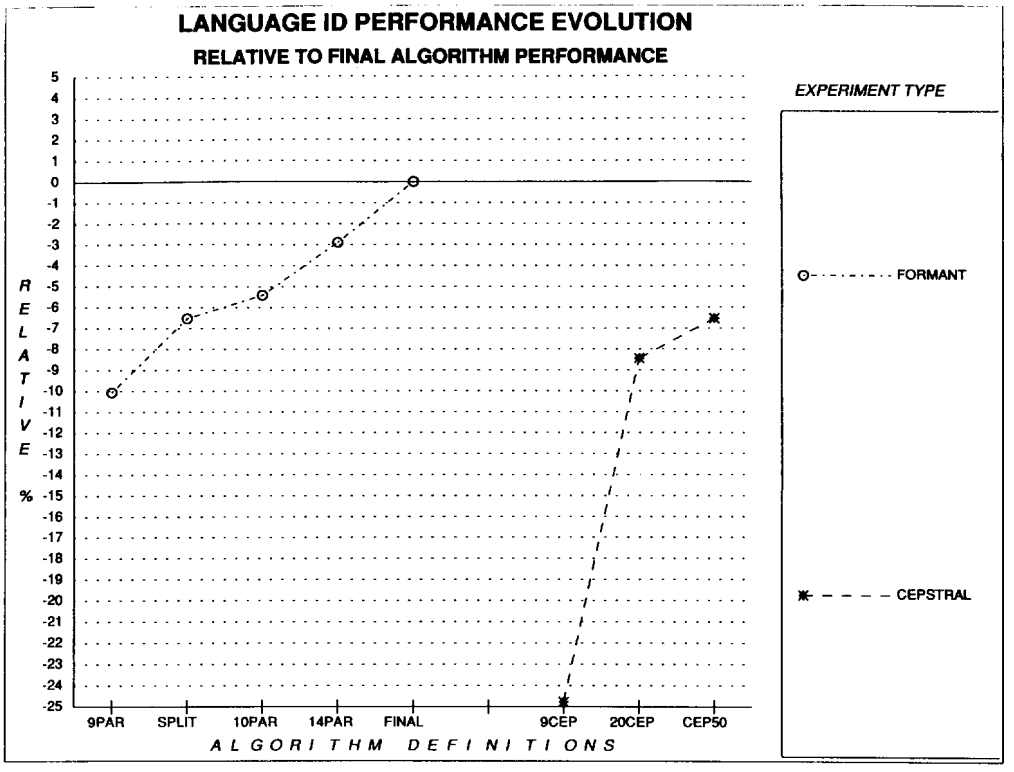

Figure 2. Algorithm Performance Evolution

FINAL TESTING RESULTS

Final testing was performed on four different language sets, including the original three-language "Foil" subset, the six-language set, and two other geographical subsets (a three-language subset, and a four-language subset). We tested variations in training, first using the training data to generate clusters, then using the original recognition data, and finally using both as training data. Though some of the sets were geographically and linguistically difficult, recognition results were superior to the earlier algorithm in all language sets. On the original set of languages, error rates were more than halved.

The performance when using the recognition clusters was very similar to when the training data clusters were used. When the the two data sets were combined, performance increased slightly. The consistent performance suggests that we were not just lucky in the selection of training vectors, or in the clustering process. Figure 3 shows the performance of various language subsets relative to the three-language geographical subset, as a function of cut length. The results show a wide variation in performance which appears to be well correlated to the "closeness" of the languages.

If no-decisions are allowed, improved recognition on the remaining cuts can be obtained through an effective no-decision rule. The rule we used was that if the difference between the top scoring languages was too small, a no-decision was made.

Recognition results were insensitive to SNR. This somewhat surprising result indicates that the formant peak-picking approach is very robust. As noise is added to a signal, the shapes of the spectral valleys change rapidly, but the peak positions change rather slowly. Thus, if the peaks are accurately chosen, this approach should outperform other methods in noise. Recognition results did show a sensitivity to spectral slope, however. A strong spectral tilt will cause an LPC front end to change the formant positions, hurting performance. This problem must be dealt with in future research.

In order to test algorithm performance on an open-set problem, several forms of rejection capability were developed. The best results were obtained when it was decided to score a "neutral" language against the input data. The neutral language (made up of samples from all six languages) is clustered the same as the others. If the neutral language scores best, all 6 languages are rejected. Otherwise, from 0 to 5 languages can be rejected and a decision is made. This approach was enhanced further by dividing the data to create a "clean" and a "noisy" set of neutral clusters. For the final test set, on average, 16 percent of the "true" languages were falsely rejected while 82 percent of the 5 "other" languages were properly rejected.

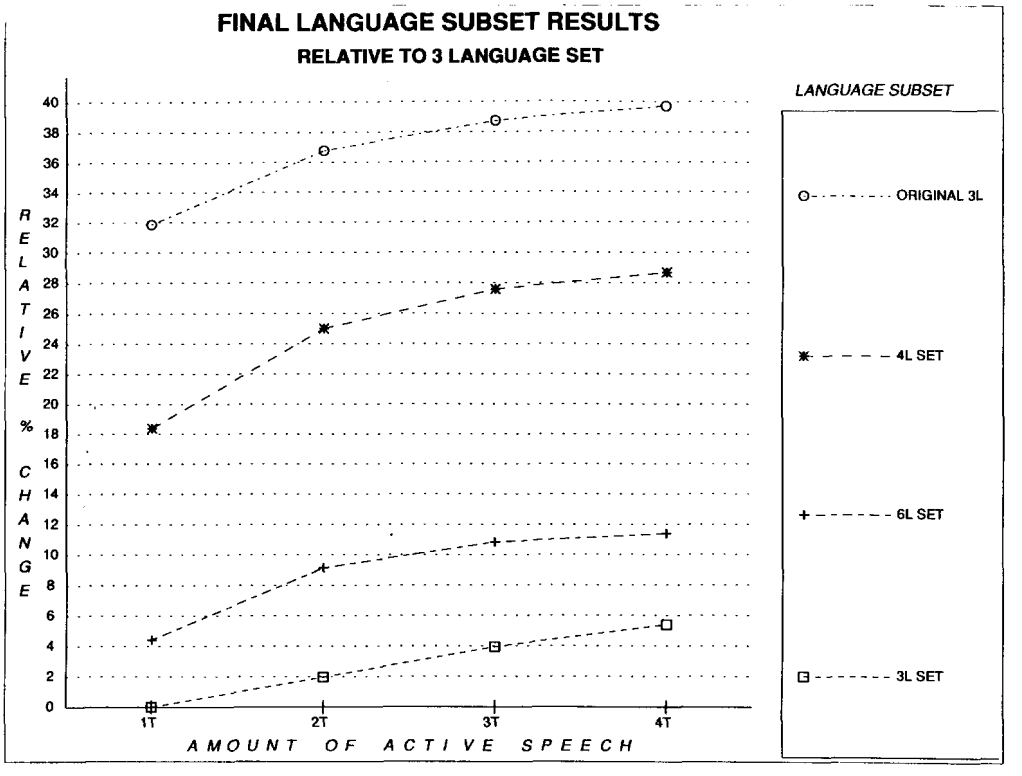

Figure 3. Relative Recognition Performance

FUTURE EFFORTS

Markov modeling is the next major area to be exploited in our research. To take advantage of its Markovian nature, the speech "states" must be defined more effectively than the current single-frame basis allows. Thus, we are converting the algorithm into a CSR (continuous speech recognizer) in which speech states are defined by a minimum and maximum dwell period, a cluster center and the Mahalanobis matrix. Once states are defined during training, a transition probabilities matrix can be developed and utilized. Variations on this approach will be tested.

We also believe that there is a great deal of potential in pitch information, in spite of our failure to exploit it. The idea of relating pitch contours to the locations of the speech states is one that deserves analysis.

SUMMARY

We have improved an existing language identification algorithm by modifying and adding parameters, improving the classifier, adding the concept of "clean" and "noisy" clusters to the training approach, and reducing its channel sensitivity. Extensive experiments on a large, noisy database confirm the improvement in the algorithm, and its robustness. A new feature of the algorithm is the successful rejection capability, giving hope that accurate open-set language ID may be possible.

REFERENCES

1. Jerry Foil, "Language Identification using Noisy Speech," Proc. ICASSP 1986, pp 861-865, Tokyo, Japan

2. J.D Markel and A.H Gray, Jr., Linear Prediction of Speech, Springer-Verlag 1976

3. Malcolm Williamson,"Gisting Analysis," Final Report Contract No. F30602-84-0001

This work was sponsored by the U.S. Air Force Systems Command, Rome Air Development Center.