Experimental Improvements of a Language Id System

Kung-Pu Li

ITT Aerospace Communication Div., San Diego, California 92131

ABSTRACT

Previously, automatic language identification systems provided good results by using syllabic "on-set" spectral features; they identified languages by finding the "nearest match" speakers who were closest to the test utterance. In this paper we show that augmenting the training data by adding speakers achieves a better gender balance in the data and reduces the error rate by more than 10%. Adding features like syllabic "coda" and "prosodic" features show very different results which can then be merged with the syllabic "on-set" spectral features to reduce errors an additional 10%. A dimensionality reduction by means of the principal components shows not only a reduction in computation and memory requirements, but also improves language identification performance when the eigenvectors are normalized with different weights. The combination of all these factors yields a significant improvement in performance when compared with the previous baseline system.

INTRODUCTION

Language identification (LID) based on speaker matching was developed with the belief that such a system could accommodate variations in speaker, dialect, and other effects within a language. The system based on this approach was presented in a previous ICASSP [1]. The basic process uses syllabic spectral features to search through (by means of a nearest neighbor algorithm) all the speakers in a multiple language database, and then, through various scoring techniques (from a set of best matched speakers), obtain the language scores for the test sample. Then, the final language decision is made. The system uses a nearest neighbor approach, applied in the language feature space as well as in the speaker space. The present study examines several important issues that affect system performance with specific databases. The goal is to optimize performance by changing several modules in the system. System performance is evaluated with respect to the number of different speakers in the reference set, the length of the reference and test utterances, errors in syllabic marking, and the use of prosodic features instead of spectral features. In the syllabic marking process, automatic training was implemented to create a new artificial neural network (ANN) to detect syllabic nuclei. This new network replaces the old network which was trained using a manual interactive process [2]. In addition, using principal components reduced dimensionality and optimized system performance by applying the proper weights to normalize the eigenvector space. This combination of factors shows consistent and accumulated improvements.

EFFECT OF THE NUMBER OF SPEAKERS IN THE REFERENCE SETS

From an error analysis of previous results [1], we suspect that the general lack of female speakers for several languages caused major errors in LID tests. Analysis of the best ten languages in a closed set identification experiment (NIST 1993 test set) shows that the average recognition rate for male speakers and female speakers differ greatly, as shown in Figure 1. This is a scatter plot for the performance of each language against the number of speakers for the language of the same gender in the reference set. The three

0-7803-2431-5/95 $4.00

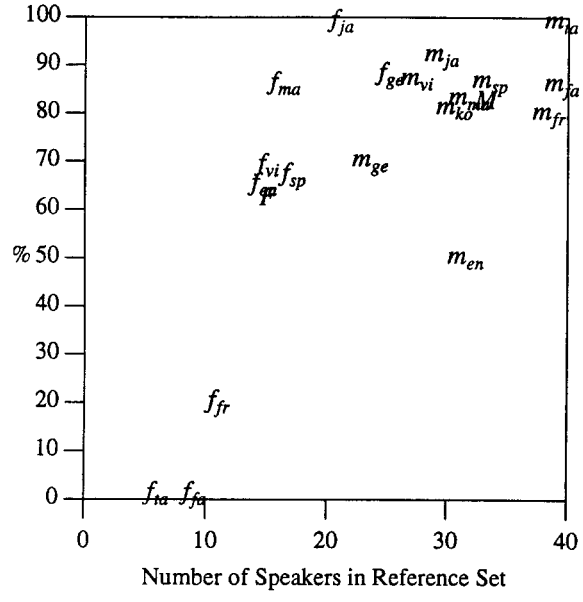

Figure 1. The scatter plot of language identification performance vs. the number of males (*m*) and females (*f*) in the reference set for each language. The average of male and female performance for all languages are shown at *M* and *F*.

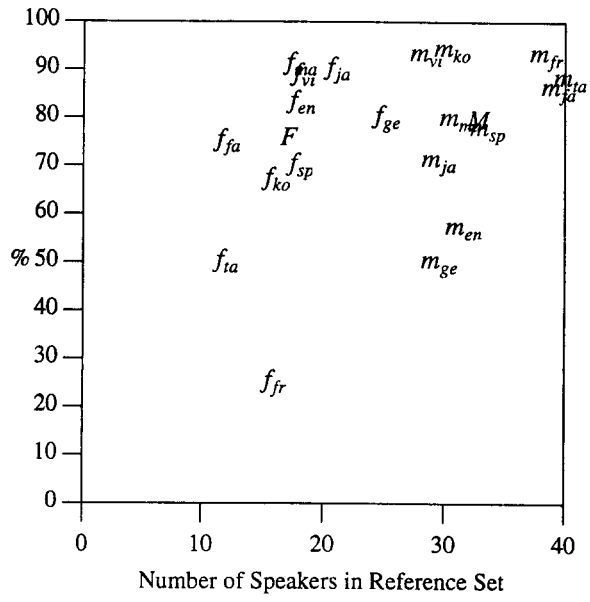

Figure 2. The scatter plot of language identification performance vs. the number of males (*m*) and females (*f*) speakers in a "balanced" reference set for each language. The average of male and female performance for all languages are shown at *M* and *F*.

languages with the least number of female speakers in their reference sets scored the worst. We can conclude that poor performance is correlated highly with the number of female speakers in the language reference set. If you ignore these three worst results (all female) shown in the left lower corner of Figure 1, you might conclude that there is no significant difference between male and female speakers.

A study tried to add more training data to "balance" the number of speakers per gender as well as the size of samples per speaker in each language. This study's test results show significant improvement. The new scatter plot is shown in Figure 2. The average performances are 79.5% (male) and 75.0% (female). (Even though few languages still lack a sufficient number of female speakers). This result indicates that the number of each gender in the reference set affects the system performance. Augmenting the training data with additional speakers to achieve a better gender balance in the reference set results in an error rate reduction of more than 10%.

SYLLABIC PROSODIC FEATURES

Normalized pitch, amplitude, and timing information (associated with the syllabic nuclei markings) are considered to be prosodic related syllabic features. For each syllabic marking, a set of features was proposed which included syllabic nuclei related pitch contours, amplitude contours, curvatures, slopes and averages, and vocalic segment duration and timing information to form a 19 dimension sample/syllable. Experiments indicate that utilizing these features from a single syllable base provides worse recognition results than combining three consecutive syllabic features (57 dimensions). Therefore, the dynamic changes in these features in adjacent syllables provides distinction among all languages. However, even with this processing, system performance is far worse than that of a system which uses syllabic spectral features, (44% vs 79% for eleven language identification). These results are obviously infe-

rior to those achieved with syllabic spectral features, as in the baseline system. However, since the performance remains above chance, some language characteristics are apparently embedded in these features. The results also indicate that significant language discrimination has little speaker variation and its error patterns are quite different from those in spectral features.

An attempt was made to merge the prosodic based and syllabic spectral feature based systems. The process combines the language scores for each testing sample. That means two systems operate independently on two sets of features for each testing sample until the process reaches the language scores. The comparisons of these test are shown in Table 1.

Test Set	Ref. Size	Num. Lang.	Syll. Spec.	Syll. Pros.	Merged Score
1993					
dev	449	10	76.0%	43.0%	78.8%
final	449	10	**73.8%**	46.4%	75.0%
1994					
final	526	10	80.4%	46.4%	**83.3%**
final	576	11	78.6%	44.4%	82.4%

Table 1. Experimental results for comparing syllabic spectral features, syllabic prosodic features and the merged system performance. The test data includes NIST-93 and NIST-94 test sets. The number of reference speakers and languages are also shown. The two results shown in bold print indicate the total performance change due to increased and balanced numbers of speaker's gender, and the system merged with the prosodic features.

This experiment uses NIST-1993 and NIST-1994 test sets for comparison. Improvements from the merging process are significant and consistent for all conditions. The reduction in error rate is about 12-15% from the baseline result. This significant improvement shows the importance of the merging effects, even when the syllabic prosodic feature yields low recognition performance (46%).

SYLLABIC "CODA" FEATURE

The improved result of merging systems using different syllabic features suggests that many different kinds of syllabic features with good language identification performance can also be merged with the system for further improvement. The test on syllabic "coda" spectral features gives a performance comparable to the syllabic "on-set" spectral features; but it encodes phonetic features on a different part of the syllabic structure. Therefore, the syllabic "coda" spectral features can be processed as an independent system, and their performance is equivalent to or slightly better than that using syllabic "on-set" spectral features (for OGI_TS 11 language database).

DIMENSIONALITY REDUCTION

Using principal components to reduce the dimensionality of the input feature space has proven to be effective, can reduce spectral feature dimensionality by more than half, and still retains more than 95% of feature variations. An optimal result also can be found by de-weighting the first six eigenvectors. The performance of language identification consistently shows better results for all three sets of features. In these experiments, we reduced the syllabic spectra features (the same as before) from 90 to 36 dimensions and syllabic prosodic features from 57 to 30 dimensions. Table 2 shows the performance comparisons with and without the eigenvector process. It also includes the results of the merged system described above.

	Syll. Fea.	Syll. Fea. w/EV
1. On-set	78.6%	81.3%
2. Coda	79.1%	83.4%
3. Prosodic	44.4%	48.1%
Merged Sys:		
1+2	83.4%	86.1%
1+3	82.4%	
1+2+3		87.7%

Table 2. The best performance results for syllabic "on-set" spectra, syllabic "coda" spectra (six frame), and prosodic features with dimensionality reduction processes. Results for several different merged systems are also shown.

This comparison indicates a significant and consistent improvement when dimensionality reduction is used. This improvement not only has the advantages of reducing memory and computation by half; it also provides better language separa-

tion. It can be speculated that the normalization of the first few eigenvector's variances increases the proportion of the language differences relative to phonetic and speaker differences in all of the tested features.

AUTOMATIC TRAINING OF ANN WITH PHONETICALLY MARKED SPEECH

In a previous paper, a neural network was trained interactively with hand-marked data from a multiple language database [2]. We used this network for the automatic detection of syllabic nuclei. In this study we develop a new training procedure for the network. It becomes fully automatic, and it has been tested on OGI-TS phonetic markings. The input to the network remains a 45-dimensional vector consisting of spectral and amplitude features from five consecutive frames. Each frame contains parameters of the energy, delta energy, the first five Olano coefficients, the right area (*i.e.*, the area to the right of the x-axis) of the circle created by plotting energy *vs.* delta energy, and the left area of the same circle. Only the frames at the center of a vowel and diphthong segments were used as target markings of syllabic nuclei. This new procedure automatically adjusts the markings during the neural network learning process. Training and development data in six languages were taken from the OGI-TS database. Data from all speakers were used, but only the first twenty seconds of speech in each speaker's file was used. To obtain a more balanced mix of positive and negative exemplars, every other negative exemplar (*i.e.*, a frame not corresponding to a syllabic nucleus) was discarded. The training set contained 163,648 exemplars; the development set contained 44,221 exemplars.

Weights were updated after presentation of each pattern in the training set using the standard back-propagation algorithm. The network was trained for 250 iterations. Twice during the training (after 100 iterations and after 150 iterations), targets were automatically adjusted as follows: if the network's output was within 1 or 2 frames from the target, the target location was adjusted to line up with the network's output. This offset any inaccuracies that resulted from the algorithm we used to mark vocalic nuclei. Given the simplicity of the marking algorithm, inaccuracy within 1-2 frames is not unreasonable. To prevent such inaccuracies from having adverse effects on subsequent learning, we adjusted the targets as described above.

At the end of training, the network could classify 92.9% of the training exemplars and 92.0% of the development exemplars correctly, where a classification is considered correct if the network's output is between 0.0 and 0.5 when the target is 0, and if the output is greater than 0.5 when the target is 1. A comparison of language identification performance using the new neural network indicates no significant changes in the overall performance for language identification; however, on further observation, we have found that the optimal performance becomes more robust and stable for the system using the new neural network.

DISCUSSIONS

This study shows that the numbers of speakers per gender in each language reference set, and the optimization in feature space with reduced dimensions are two important factors for speaker-based language identification. Furthermore, merging systems that use different front-end features can provide more consistent and better results than any single system. The study also created an automatic training procedure to obtain ANN weights with phonetic marking. From those improvement, the performance of a specific database (OGI-TS) increased from 79% to 88%.

REFERENCES

[1]. Li, K. P. "Automatic Language Identification Using Syllabic Spectral Features", ICASSP-94, pp. I-297-300, 1994.

[2]. Li, K. P. "Neural Network Approach to Assist Markings of Syllabic Nuclei for Multi-language Database", JASA 92-4(2) pp. A-2477, 1992.