LANGUAGE IDENTIFICATION USING NOISY SPEECH

Jerry T. Foil

GTE Government Systems Corporation*
Mountain View, California, USA 94042

ABSTRACT

This paper describes experiments in automatic identification of spoken languages using recordings of noisy radio signals as a data base. Prior efforts used uncorrupted speech; we selected techniques that we believed would be robust in noise. One technique attempted to distinguish languages by applying a classical quadratic classifier to prosodic features extracted from pitch and energy contours. Another was designed to exploit the frequency of occurrence of characteristic sounds using formant locations to represent the sounds, and using a vector-quantization distortion measure as the basis for language decisions. The techniques were required to make decisions based on speech segments of a few seconds duration. Our final tests were conducted on over 4 hours of previously unprocessed speech. Three languages, each from a different major language group, were used for development and testing. Allowing 11% false rejection (no decision), we achieved 64% correct identification with short speech segments. Our plans include the application of Markov modeling techniques to language identification.

INTRODUCTION

Automatic identification of languages is a difficult problem that has been the subject of but a few past research efforts. One approach was to use word-spotting techniques to identify multi-phoneme, reference sounds [1]. Another applied advanced pattern recognition techniques to frame-by-frame speech parameters [2]. Both efforts used uncorrupted speech and based decisions on speech segments with durations on the order of 2 minutes. Many operational applications of language identification require that identification be performed on poor quality signals within a few seconds.

The objective of our work was to develop and test techniques that are robust to noise and that are capable of providing decisions in 10 seconds or less. The specified approach included the following:

a. Analysis of operational tapes containing signals representative of those against which the algorithm must operate.

b. Development of multiple, candidate, language-identification techniques.

c. Optimization of these techniques using independent training and test data taken from the operational tapes.

d. Final testing using previously unprocessed tapes.

DESCRIPTION OF TECHNIQUES

Processing Pitch and Energy Contours

The field of contrastive linguistics attempts to formally describe differences among languages. A study of linguistics literature brought us upon the idea that prosodic features could be the basis of a powerful language identification technique [3,4]. It is obvious that rhythm and intonation patterns vary from language to language. This prosodic information should be contained in pitch and energy contours. To test the validity of our hypothesis, two experiments were conducted:

1. An experienced linguist examined plots of pitch and energy contours from two languages (one Slavic and one tonal south-east Asian). The linguist was able to distinguish between the languages based solely on these plots.

2. Using LPC analysis-synthesis, we removed the vocal tract information from speech of each language (by holding the filter coefficients constant during synthesis). The result gave the impression that the speaker's mouth was closed. The only information remaining in the signal was that of pitch and energy. Although the speech was unintelligible, it was quite easy to distinguish between the languages.

In both experiments, the amount of speech required to distinguish the languages was less than 10 seconds, and the speech used was from our operational data base.

To exploit prosodic differences among languages we applied feature-based pattern recognition techniques. Based on interviews with linguists, we developed a list of 45 candidate features that could be extracted from pitch and energy contours. We also developed an endpoint detection algorithm designed to identify beginnings and ends of utterances. The objective was for an utterance to roughly correspond to a sentence. Table 1 lists the major attributes of the contour-endpoint detection algorithm. Table 2 summarizes the 45 candidate features.

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

 CH2243-4/86/0000-0861 $1.00

Table 1. Contour-Endpoint Detection Logic

Measurements used: 1. Height of cepstral pitch peak 2. Current energy compared to noise-riding threshold 3. Zero-crossing rate
Decision technique: 1. Each frame is classified speech/nonspeech using a quadratic classifier. 2. Frame-by-frame decisions are smoothed and subjected to decision logic
Training data: 1. One hour of speech 2. Taken from operational tapes

Table 2. Summary of Features Examined

Feature Number	Energy
1	Length of contour
2-7	Features based on distances between peaks in energy contour
8-13	Features based on the derivative of the energy contour
5-17	Normalized central moments of histograms of frame-by-frame energy
18-22	Features based on widths of peaks in energy contour
23	Correlation coefficient between energy and pitch contours
	Pitch
24-29	Features based on pitch slope (magnitudes and relative locations)
30	Average pitch
31-35	Features based on pitch fluctuation
36-40	Features based on extremes of contour
41	Percent of contour from voiced speech
42-45	Normalized central moments of histograms of pitch values

The fact that the feature set contained some redundancy was not a concern. The 45 features were a superset from which an "optimum" subset would be selected. The subset was selected using a 2,1 heuristic feature search (HFS) algorithm.

Given a chosen classifier, a set of training data, and a set of test data, the 2,1 HFS algorithm searches for the feature subset that minimizes a user-defined probability of error by alternately selecting the best two features and discarding the worst one feature until the desired feature-set size is reached. For this feature search, we selected a quadratic classifier (others were tried later), and used 6 hours of speech from each of three languages as a data base. The data base was divided evenly into training and test sets.

This process resulted in the selection of seven features. Addition of the eighth-best feature did not improve performance. Table 3 lists the seven features and gives comments on what those features measure.

Table 3. List of Chosen Features

Feature Number	Feature	Comments
1	Number of frames in contour	Length of contour corresponds to length of utterance
2	Normalized second central moment of frame-by-frame energy values	Shapes of histogram of energy values and energy derivative values tended to vary among the languages.
3	Normalized third central moment of frame-by-frame energy values	Normalized central moments characterize histogram shape.
4	Normalized third central moment of frame-by-frame energy derivative values.	
5	Average of the absolute value of frame-by-frame change in energy (normalized by average energy).	Measures relative high-frequency content in energy contour.
6	Maximum pitch deviation	Measures pitch range over short periods.
7	Percent voiced speech in contour	Tonal languages have more and longer voiced segments.

The HFS was repeated using a nearest-neighbor classifier and a nearest-cluster classifier. Performance with the quadratic classifier had been the best. The quadratic classifier is optimum if the features are distributed in a joint-Gaussian fashion. Histograms and scatter-plots of the features indicated that although they were not perfectly Gaussian, the distributions were unimodal and fairly symmetric, giving no reason to believe that the quadratic classifier was a poor choice.

Processing Clusters in Formant-Space

The phonemes that make up one language often differ from those that make up another. Even in languages with very similar phoneme sets, the frequency distributions of the phonemes often vary between the languages. Phoneme recognition, or phoneme-type recognition, could form the basis of a powerful language-identification technique. Unfortunately, the signals in our data base were so noisy that explicit phoneme recognition was not possible. We sought to find an automated method of exploiting the differences in phonemes between languages, without having to make a phoneme recognition decision.

If a parameterization of the vocal tract transfer function (such as LPC coefficients, a filter bank output, or formant data) is treated as an abstract vector, one

would expect that if a large set of vectors was collected from unconstrained speech the common sounds of the language would be seen as clusters in the space spanned by the speech parameters. If the common sounds of one language differ from those of another, one would expect the cluster locations to vary between languages. This assumption was the basis of our second technique. Since clusters could be found without knowing the sounds they represent, explicit phoneme recognition could be avoided.

Robustness drove the choice of parameterization. Formant values were chosen for two key reasons: (1) It is known that the human ear and brain make heavy use of formants to distinguish sounds, and (2) additive wideband noise has the least effect on the peaks in the spectrum.

The algorithm used to process formant vectors is outlined in Table 4. Using 2 hours of speech from each of three languages, formant vectors were collected from all voiced speech frames. For each language the vectors were clustered using a k-means clustering algorithm. The k-means algorithm has the following property: Given a set of vectors to be clustered, a definition of a distortion measure, and the number of clusters sought, the k-means method iteratively determines the clusters that minimize the average distortion between each vector and the center of its cluster [5].

Table 4. Outline of Formant-processing Algorithm

I.	Off-line training.	
	A.	Formant vectors from voiced frames of 6 hours of speech are collected.
	B.	For each of the three languages, a k-means clustering algorithm finds the best 10 clusters.
II.	Decision algorithm.	
	A.	If the input frame is voiced, the formant values are calculated.
	B.	The formant vector is compared to each cluster-center from each language.
	C.	The closest cluster-center for each language is found, and the distance to that cluster-center is noted.
	D.	The noted distance for each language is accumulated across all voiced frames.
	E.	The language with the smallest accumulated distance is the most likely.

The decision algorithm uses the locations of the cluster centers. Given a formant vector from speech to be classified, the algorithm locates the nearest cluster center for each language and stores the distance to that cluster center. As more formant vectors are processed, the distances for each language are accumulated. The most likely language is then the one with the lowest accumulated distance. This is equivalent to basing the decision on a vector-quantization distortion measure.

RESULTS

Two phases of testing were conducted. They are referred to here as the development tests and the final tests. The purpose of the development tests was to provide feedback for the algorithm development process. This allowed selection of the best features for the pitch-and-energy algorithm, and selection of the best number of cluster centers to use in the formant-cluster algorithm. It would not be appropriate, however, to report the results achieved during development because the algorithms were optimized for that set of test data. Final tests were conducted using 2.5 hours of previously unprocessed speech from each language. The only parameters allowed to vary during final tests were the acceptance-rejection thresholds.

Pitch-and-Energy Algorithm

During development the pitch-and-energy algorithm was achieving performance percentages in the low seventies using an average of 5 seconds of speech to make a forced decision among three languages. During final test, the performance dropped to 39%. One explanation for the disappointing drop in performance is the fact that there are over 45 million combinations of 6 features that can be chosen from a list of 45. It may be that the six features chosen were the "luckiest" set given the training-test data partition, while they actually had very little inherent language discrimination ability.

Formant-Clustering Algorithm

The formant-clustering algorithm performed well during both development and final tests. Using an average of 5 seconds of speech to make a forced decision among three languages, the algorithm achieved 69% correct decisions during development. During final tests a no-decision category was added to the decision logic. Allowing no decision to be made on 11% of the inputs, 64% were classified correctly. For this test the average signal duration was 4.5 seconds, and the typical signal-to-noise ratio was 5 dB.

The decisions were made using the following logic: If the lowest accumulated distance is less than a given absolute threshold, and is exceeded by the accumulated distances from the other languages by a sufficient amount, then a decision is made. Otherwise another formant vector is processed and the logic repeated. If the end of a speech segment is reached and no decision has been made, the input is not classified.

Another experiment was conducted to determine the ability of the algorithm to make an explicit language rejection decision, that is to decide that the input is not one or more of the candidate languages without necessarily deciding which language it is. Allowing 6% false rejection, 18% of the inputs were correctly rejected.

NEW EFFORTS

To improve language-identification performance, we plan to focus our efforts in the following areas: (1) A more sophisticated approach to pitch-and-energy processing will be tried. Our experiments showed that pitch and energy contours have the information needed to discriminate languages, but our algorithm failed to exploit that

information. (2) Recent developments in robust formant trackers will be investigated. (3) Markov modeling will be tested.

SUMMARY

We have reported the results of experiments in language identification using short segments of noisy speech as a data base. We selected techniques that had the potential to capture language information in these signals. One of the techniques, based on clusters of formant vectors, achieved 64% correct identification, on signals of 4.5 seconds duration and an average signal-to-noise ratio 5 dB.

REFERENCES

[1] Dr. Gary L. Leonard, "Language Recognition Test and Evaluation," RADC-TR-80-83, 1980

[2] D. Cimarusti, and R.B. Ives, "Development of an Automatic Identification System of Spoken Languages: Phase I," Speech Communication Research Laboratory, 1981

[3] L.R. Waugh, and C.H. Van Schooneveld, The Melody of Language--Intonation and Prosody, University Park Press, 1980.

[4] A. Cutler, and D.R. Ladd, Prosody: Models and Measurements, Springer-Verlag, 1983.

[5] M.R. Anderberg, Cluster Analysis for Applications, pp 160-163, Academic Press Inc., 1973.