Evolutionary Programming for Voice Feature Analysis

D.B. Fogel L.J. Fogel

ORINCON Corporation
9363 Towne Centre Drive
San Diego, CA 92121

ABSTRACT

The problem of using speech features in polygraphy is addressed. Two techniques are analyzed. Fisher-optimal discriminant features are used with insignificant results. Evolutionary Programming is used to evolve suitable classification models. In two experiments involving multiple subjects, the evolutionary approach was able to achieve about 70% correct classifications, while examining only 10^{-7} of all possible feature combinations.

Introduction

Previous work [1] centered on a mock theft experiment involving ten adult subjects (five male, five female) in which infrared (IR) and speech measurements were made in conjunction with three standard polygraph channels (pneumograph, pressure cuff, and galvanic skin response). The object of the experiment was to test the efficacy of speech and IR for lie detection, to determine which features (if any) of the speech and IR are important for polygraphy, and to relate these features to standard polygraph measurement.

The results of the series of experiments were encouraging. Using four speech features, the average classifier yielded perfect classification ($P<0.05$) of all suitable test subjects. IR classifiers were also able to discriminate between the test subjects. The nine resulting IR classifiers, based on a majority vote of five univariate classifiers, performed perfectly. The univariate classifiers themselves decided correctly on 44 of 45 trials.

The current work focused on establishing whether or not there is sufficient information in voice patterns to facilitate accurate and reliable real world polygraphy in real world situations. Accuracy was defined by the ability of the speech classification system to discriminate between "guilty" and "innocent" responses. Reliability was defined as a consistency of classifying features across trials.

Analog recordings using Kyocera D-611 cassette recorders with Audio Technica AT813 unidirectional condenser micropnones of 15 actual interviewee responses (five known "stressed," five known "unstressed," five unknown) of the word "no" from each of 12 subjects were received from the sponsor. These responses were digitized on the Macintosh using an 8-bit digitizer sampled at 22 KHz. Each utterance contained approximately 5,000 samples. The gain on the digitization was adjusted for each response to make the maximum voltage roughly equal for each response. Equal volume level over responses is required for the speech feature extraction. This precludes the use of response volume as a discriminating feature for use in classification.

The speech feature extraction program requires the following user inputs: initial detection threshold — the signal level above which onset of the utterance is detected; glottal pulse threshold — the signal level above which a glottal pulse is detected; and pulse peak search window — the sample length of the window from a glottal pulse peak to the next sample considered to be in a separate glottal pulse. Establishing consistent thresholds for all responses was the reason for making the approximate volume level for all responses the same. The speech feature extraction program was executed with the same threshold parameter settings for all responses. 101 features were produced for each response.

Discussion and Experimental Procedures

The 15 response vectors from each of ten of the twelve subjects (101 real numbers per response) were then run through the classifier. This classifier first normalizes each response component to have a mean of zero and unit variance. All of the responses, training and test, were used to compute normalization parameters. Optimal discriminating features were selected based on the Fisher statistic:

$$F = \frac{|\bar{X}_s - \bar{X}_u|}{\sigma^2_s + \sigma^2_u}$$

where $\bar{X}_s$ and $\bar{X}_u$ are the average feature values in the stressed and unstressed training samples respectively, and σ^2_s and σ^2_u are the respective variances of the feature values in the stressed and unstressed training samples. This value was computed for all 101 response components. The "best" feature components for classification were chosen to be those with the highest Fisher statistic. Classifications were made using the best one and two features for classification.

23ACSSC-12/89/0381 $1.00

The results for this classification are indicated in Table I. Each subject was analyzed separately. When a two feature analysis was conducted, the features were chosen independently; they are not the best paired features. Overall, using only the best feature, 26 of 49 classifications were correct. Using the best two features, 24 of 49 classifications were correct. These results do not provide sufficient evidence to reject a null hypothesis that these results were due simply to chance.

In light of the insignificant results obtained using the standard classifier technique, an alternative method was sought which would discover the best multiple-feature classification of the subjects in speaker independent and dependent analysis. Evolutionary Programming, conceived by Fogel, [2] was recognized as an optimization technique which has been used to find optimal parameter settings in difficult multi-dimensional optimization problems.

Table I
Classification Using Fisher-Optimal Classifiers

		One Feature		Two Features	
Subject #	Actual	(# correct)		(# correct)	
1	igggg	giggi	(2)	giggg	(3)
2	gggig	gigig	(4)	iiggg	(2)
3	giggg	giigi	(3)	giigi	(3)
4*	iiigi	iii?i	(4)	gig?g	(1)
5	ggigi	iiiig	(1)	iiigg	(2)
6	ggigg	iiiig	(2)	giiig	(3)
7	igggg	giggi	(2)	giggi	(2)
8	giiii	iigig	(2)	iigig	(2)
9	iiigg	giigg	(4)	giigg	(4)
10	ggiii	iigii	(2)	iigii	(2)

i = "innocent" response
g = "guilty" response
? = unable to classify.
*It was not possible to digitize the fourth response for subject #4

The original Evolutionary Programming concept focused on the problems of predicting any stationary or nonstationary time series, modeling an unknown transducer on the basis of input/output data, and optimally controlling an unknown, all plant with respect to an arbitrary payoff function [3,4]. Natural evolution optimizes behavior through iterative mutation and selection within a class of organisms. Behavior can be described in terms of the stimulus/response pairs that depend on the state of the organism. Each organism can be portrayed as a finite state machine, a mathematical function that does not constrain the represented transduction to be linear, passive, or without hysteresis.

The evolutionary process is simulated in the following manner: an original population of "machines" (arbitrarily chosen or given as "hints") are measured in their ability to predict each next event in their "experience" with respect to whatever payoff function has been prescribed. Progeny are then created through random mutation of these "parent" machines. These offspring are scored on their predictive ability in a similar manner to their parents. Those organisms which are most suitable for achieving the task at hand are probabilistically selected to become the new parents. An actual prediction is made when the predictive-fit score demonstrates that a sufficient level of credibility has been achieved. The surviving machines generate a prediction, indicate the logic of this prediction, and become the progenitors for the next sequence of progeny, this in preparation for the next prediction. Thus, aspects of randomness are selectively incorporated into the surviving logics. The sequence of predictor machines demonstrates phyletic learning, through the inductive generation of hypotheses concerning the relevant regularities found within the experienced environment in light of the given payoff function.

Recently, Evolutionary Programming has been used for the optimization of difficult combinatorial problems. In traveling salesman problem experiments where 100 cities were randomly distributed in a square area in accordance with a uniform distribution, results indicate that the evolutionary technique discovered final routings better than 99.99999999999% of all possible tours, while examining only 8.58×10^{-151} of all possible cases [5,6]. Discovering the optimal set of features for the classification of each subject's response is a combin-atorial optimization problem. The task is to choose which set of features, from the 101 possible features, provides the most accurate classification, while minimizing the total number of chosen features.

The typical Evolutionary Programming procedure was adapted in the following manner. Rather than evolve finite state machines, the organisms were represented by bit strings of length 101. Each bit corresponded to a given feature of speech. If a specific bit were "on," that corresponding feature was indicated to be included in the model; if the bit were "off," the feature was excluded. The total number of possible models is 2^{101}. Because of the sample size limitations, and in light of earlier work, the maximum number of features that would be allowed in any model was set to five. Thus, there were slightly more than 9.6×10^9 possible contending models.

Each evolving model was scored in its ability to perform classification on a training set, weighted by the number of features used in the model. A jackknife procedure was used. The training data were composed of the known stressed and unstressed responses from individual or multiple subjects. It was required that the model classify each conditioned response, based on the remaining known responses. Classification was made as to whether the normalized conditioned response was closer to the mean of the normalized stressed or normalized unstressed group. No responses were allowed to be "unclassified." Akaike's Information Criterion [7] was used to determine an overall worth for each evolving model.

Mutation was conducted by randomly choosing from one to five bits in each parent model and reversing their signs, subject to a maximum of five features being indicated in the offspring model. Twenty-five parent models were maintained at each generation. After examining 2,500 models in single subject experiments, and 5,000 models in multiple subject

experiments, each unknown response was classified using the best evolved model.

Experiments were conducted on a speaker dependent (individual subject) and a speaker independent (multiple subjects) basis. The results are indicated in Table II. While the results of the speaker dependent experiments are widely varied, the evolutionary technique was able to correctly classify close to 70% of the responses in both speaker independent experiments, while examining about 10^{-7} of all possible models.

Table II
Classification Using Evolved Classifiers

Subject #	Actual	Sing. Subj. (# correct)		Grouped Subj.* (# correct)	
1	igggg	giggg	(3)	ggggg	(4)
2	gggig	giigg	(2)	gigig	(4)
3	giggg	ggggg	(4)	iiiii	(1)
4**	iiigi	gii?g	(2)	iig?i	(3)
5	ggigi	ggigg	(4)	ggigi	(5)
6	ggigg	iiiii	(1)	igigg	(4)
7	igggg	iiggg	(4)	ggiig	(2)
8	giiii	iiggg	(1)	iiiii	(4)
9	iiigg	iiiig	(4)	iiigg	(5)
10	ggiii	ggggg	(2)	iigii	(2)

i = "innocent" response
g = "guilty" response
? = unable to classify.
*Subjects 1-5 were grouped, and subject 6-10 were grouped
**It was not possible to digitize the fourth response for subject #4

Conclusions

Evolutionary Programming is much more effective in discovering an optimal set of features to perform the classification task than the use of one or two Fisher-optimal features. The variability of performance on single subject classifications is probably due to small sample sizes. To reduce variability it would be of interest to collect more samples per subject and then reexamine the per-formance of the algorithm, perhaps combining it with standard techniques in polygraphy (*e.g.*, galvanic skin response, pneumograph, pressure cuff).

While there was very little consistency in speech features across subjects, the need for similar features in practice is reduced using an evolutionary approach. The parallel nature of the algorithm makes its suitable for distributed processing machines. Given an appropriate architecture, a suitable set of features could be quickly discovered, easing the need for fixing the discriminating speech features *a priori*.

In addition to a gain in performance, Evolutionary Programming offers the advantage of payoff functions that are significantly different from all-or-none. Depending on the circumstances, it may be more important to correctly identify stressed responses than to identify unstressed responses (*e.g.*, airport security); the possible Type I and Type II errors may also not be of equal cost.

This work should encourage further investigations including larger sample sizes, voices modified to simulate telephone transmission, and an increased vocabulary of possible responses.

References

[1] R. Altes, "Infrared and Speech Processing for Polygraphy," Final report under contract #83F802500, December 1985.

[2] L.J. Fogel, "Autonomous Automata," Industrial Research, vol. 4, pp. 14–19, 1962.

[3] L.J. Fogel, On the Organization of Intellect, Ph.D. Dissertation, UCLA, 1964.

[4] L.J. Fogel, A.J. Owens, M.J. Walsh, Artificial Intelligence Through Simulated Evolution, (John Wiley & Sons, New York), 1966.

[5] D.B. Fogel, "An Evolutionary Approach to the Traveling Salesman Problem," Biological Cybernetics, vol. 60, pp. 139–44, 1988.

[6] D.B. Fogel and L.J. Fogel, "Route Optimization Through Evolutionary Programming," Proc. 22nd Asilomar Conf. on Signals, Systems, and Computers, October 1988.

[7] H. Akaike, "A New Look at Statistical Model Identification," IEEE Auto Control, vol. 19, no. 6, 1974.