2004 IEEE 11th Digital Signal Processing Workshop
& IEEE Signal Processing Education Workshop

# EVALUATION OF SPHERICALLY INVARIANT RANDOM PROCESS PARAMETERS AS DISCRIMINATORS FOR SPEAKER VERIFICATION

*Joseph San Filippo[1] and Phillip DeLeon[2]*

[1]Honeywell Technical Solutions Inc.
NASA White Sands Test Facility
Propulsion Test Dept.
Las Cruces, NM 88004-0020
sanfilippo@zianet.com

[2]New Mexico State University
Klipsch School of Elect. & Comp. Eng.
Box 30001 / Dept. 3-0
Las Cruces, New Mexico USA 88003
pdeleon@nmsu.edu

## ABSTRACT

Current methods of speaker identification and verification rely on the complex extraction of hundreds or even thousands of parameters in order to correctly model and identify a speaker. These methods have matured to the point where extremely accurate identification of a speaker (from a large population of speakers) is possible. In this work, we are interested in the potential use of Spherically Invariant Random Processes (SIRPs), described by two parameters, for speaker identification. These random processes have been shown to be a more statistically-accurate model for speech than Laplace and Gamma pdfs. Computation of the two SIRP parameters is fast and simple and storage requirements are obviously small. Although the proposed method does not yield the accuracy of current methods, identification rates are better than random guessing. The work demonstrates the first step for potential use of SIRPs in speaker identification. Usage might include an adjunct role where SIRPs could supplement existing methods to further improve identification or be used to reduce the parameter requirements of existing methods while maintaining accuracy rates.

## 1. INTRODUCTION

Current methods of speaker identification and verification rely on the complex extraction of hundreds or even thousands of parameters in order to correctly model and identify a speaker. These methods have matured to the point where extremely accurate identification of a speaker (from a large population of speakers) or accurate verification of a speaker is possible [1], [2], [3]. As an example, for the speaker identification technique described in [2], 20 mel-cepstrum coefficients (feature vectors) are extracted every 10ms for each training utterance (90s in length), translating to 180,000 feature parameters. These parameters are then modeled with a Gaussian Mixture Model (GMM) (20 mixtures) yielding 60 discriminators (20 weights, 20 means, and 20 variances) per speaker. The approach is computationally complex (although all computation is performed prior to identification) but extremely accurate.

We consider a speech signal to be a realization of a random process, that process being embodied by the particular individual producing the speech. Random processes are typically characterized by such properties as probability density functions (pdfs), correlations, and moments. In this paper, we characterize the training utterance as a Spherically Invariant Random Process (SIRP) simply described with two parameters [4]. Our interest is in evaluating the potential of an extremely-low number of discriminators in speaker verification (SV).

In work by Brehm and Stammler, it has been demonstrated that SIRPs can give rise to first order pdfs that provide a more statistically accurate model for band-limited speech than Laplace and Gamma pdfs (which are in fact special cases of SIRPs) [4]. The authors have developed a tractable method for computing the first order pdfs for these SIRPs which are completely described by a pair of parameters referred to as $b_1$ and $b_2$. Subsequent work suggested that each speaker in a limited set may be uniquely characterized by the ($b_1$, $b_2$) pair that generates the closest-fit pdf to the empirically determined histogram of the speech signal [5].

In this work, we show the results of utilizing the SIRP parameters in speaker verification. This paper is organized as follows. In Section 2 we offer a brief review on SIRPs and their description using *G*-functions. This description includes the two parameters, $b_1$ and $b_2$. In Section 3 we discuss the YOHO speech corpus used in the experiments and in Section 4, discuss the use of a parameter in addition to the two SIRP parameters for better discrimination. In Section 5 we detail the experiments and discuss the results. Finally, we conclude the paper.

0-7803-8434-2/04/$20.00©2004 IEEE

## 2. SPHERICALLY INVARIANT RANDOM PROCESSES AND MEIJER'S *G*-FUNCTION

Empirical studies of telephone-band-limited speech show bivariate PDFs with elliptical contour lines of equal height for time differences not exceeding 5 ms [4]. Hence, spherically invariant random processes (SIRPs), which are characterized by bivariate PDFs with elliptical or circular contour lines, were introduced to model band-limited speech [4]. SIRPs are shown to better model empirically determined univariate speech PDFs than the traditionally used Gamma, Laplace, and $K_0$ densities [4]. It is of interest to note that SIRPs, unlike most random processes in general, are completely characterized by their univariate PDF and correlation function [4].

Brehm and Stammler show that Meijer's *G*-function can be used to compactly express univariate SIRP PDFs; that the Laplace, $K_0$ , and Gamma PDFs are SIRPs; and that the associated univariate densities are members of the family of *G*-functions represented as:

$$p(x) = A\,G_{0\ 2}^{2\ 0}\left(x^2 \middle| b_1, b_2\right). \qquad (1)$$

This function is characterized by two interchangeable parameters, $b_1$ and $b_2$. The *G*-function is a generalization of a hypergeometric function and is described in terms of the Melin-Barnes integral (for more details see [4]). For our purposes, we need only focus on the *G*-function parameters. Table 1 shows values of $b_1$ and $b_2$ for the above-mentioned densities. By varying $b_1$ and $b_2$, a variety of improved modeling functions can be generated and a better statistical fit to actual speech signals can be achieved [4].

Table 1: *G*-function parameters, $b_1$, $b_2$ for some common probability density functions

| | $b_1$ | $b_2$ |
|---|---|---|
| Laplace | 0.0 | 0.5 |
| Gamma | -0.25 | 0.25 |
| K0 | 0 | 0 |

## 3. THE YOHO VOICE VERIFICATION CORPUS AND SIRP CALCULATIONS

The YOHO voice verification corpus is a standard database for testing speaker identification and verification systems [7]. It is available from the Linguistic Data Consortium. The corpus consists of "combination lock" phrases spoken by 138 individuals in an office environment. There are 4 enrollment sessions per speaker, with 24 utterances per session (only 3 of the 4 sessions were used). There are10 verification sessions per speaker, with 4 utterances per session.

The general experimental procedure was: for each speaker calculate PDFs for enrollment and verification sessions using amplitude histograms of speech samples; for each PDF so calculated, find $b_1$ and $b_2$ such that (1) yields a close match to the empirical PDF; assess the performance of the resulting $b_1$, $b_2$ pairs for correctly matching enrollment speakers to verification speakers. A random search algorithm was used to generate $b_1$, $b_2$ pairs assumed to provide the closest match to the empirical PDFs. For each search, 4000 PDFs were calculated and compared to the histogram; the PDF with the minimum mean square error relative to the histogram was taken to be the closest-match. Calculation of 4000 PDFs required a few minutes of computation time in MATLAB.

## 4. SPECTRAL DISCRIMINATORS, FEATURE VECTORS, DISTANCE MEASURES

Experiments were conducted to determine if the addition of a third, relatively simple, parameter to each ($b_1$, $b_2$) pair would provide improved discriminating power, i.e. separation of speakers. Various spectral properties of the utterances were considered. For purposes of these experiments, the third discriminator was defined to be the fraction of the total energy in the spectrum that occurs above 1 kHz. [Combining this fraction with ($b_1$, $b_2$) was found to be convenient because the fraction was seen to be generally of the same order of magnitude as $b_1$ and $b_2$.]

For enrollment, a $b_1$, $b_2$ pair was calculated for each of 72 utterances per speaker, then a single average $b_1$ and average $b_2$ were calculated. In addition, the average over the 72 utterances of the fraction of total energy above 1 kHz was calculated. This ($b_1$, $b_2$ , energy) triplet formed a 3-element feature vector intended to characterize each speaker.

For verification, a single average $b_1$ and average $b_2$ were calculated for each session, resulting in 10 $b_1$, $b_2$ pairs per speaker. In addition the average over the 4 utterances in each session of the energy above 1 kHz was calculated. This resulted in a single ($b_1$, $b_2$ , energy) triplet for each verification session, allowing 10 tests per speaker.

The measure of "closeness" between a particular verification feature vector and a given enrollment feature vector was the simple Euclidian distance between the 2 vectors in ($b_1$, $b_2$ , energy) space.

## 5. SPEAKER VERIFICATION EXPERIMENT

Experiments were performed to evaluate the performance of SIRPs, supplemented by the third spectral discriminator, in a text-independent SV application, two experiments were performed, one to determine a false

rejection (FR) rate and one to determine a false acceptance (FA) rate.

The FR experiment was performed for all verification sessions representing claimants, under the condition that all claims are correct, i.e. a set of verification utterances from speaker $n$ are assumed to be accompanied by the correct claim that the speaker is speaker $n$. We test to determine if the system accepts or falsely rejects the claim.

### 5.1 FR Experimental Procedure

The methodology for the FR experiment was as follows for a verification session known to belong to speaker $n$:

Step 1: Obtain the ($b_1$, $b_2$, spectral discriminator) vector for the verification session.

Step 2: Calculate the vector distance between this verification vector and each of 138 speakers' enrollment vectors.

Step 3: Count the number of enrollment speakers, $N$, whose enrollment vectors are farther from speaker $n$'s verification vector than is speaker $n$'s enrollment vector.

Step 4: Define a threshold. If $N$ exceeds the threshold, we consider that the claim would be accepted, i.e. we would accept the claim that a speaker producing the tested verification session is speaker $n$. If $N$ is less than or equal to the threshold, we would reject the claim.

The above test was performed for all speakers as claimants and false rejection rates were determined for a range of thresholds.

### 5.2 FA Experimental Procedure

The FA experiment was performed for all verification sessions representing claimants, under the condition that all claims are incorrect, i.e. a set of verification utterances from speaker $n$ is considered to be accompanied by the false claim that the speaker is speaker $p$. We test to determine if the system would reject or falsely accept the claim. For each verification session, the falsely claimed speaker was determined randomly.

The methodology for the FA experiment was essentially the same as for the FR experiment, except that for verification speaker $n$ and falsely claimed enrollment speaker $p$, speaker $p$'s enrollment session is substituted for speaker $n$'s enrollment session.

To determine whether to falsely accept the claim, we count the number of enrollment speakers, $N$, whose enrollment sessions are farther from speaker $n$'s verification session than is speaker $p$'s enrollment session. We are, in effect, determining if speaker $n$ is a good impostor for speaker $p$. If speaker $n$ is a good impostor, we will falsely accept the claim that speaker $n$ is speaker $p$. FA rates were determined for a range of thresholds.

### 5.3 The Effects of Threshold on Performance

Any biometric means of verification or identification involves imperfect testing. In the general case of verification, a test is performed and the outcome of the test must exceed some threshold in order for the claim to be accepted; there will always be false rejections and false acceptances. In our SV experiment, the number, $N$, of speakers whose enrollment sessions are more distant from the claimant's verification session than the claimed speaker's enrollment session was required to exceed a threshold. For a simple numeric threshold such as this, the value of the threshold will directly affect system performance. This is analogous to setting a bar over which a claimant (either a "client," i.e. a valid claimant, or an "impostor," a claimant making a false claim) must jump in order to be accepted. Set the bar too low ("too easy") and the number of impostors admitted (false acceptance) will increase. Set the bar too high ("too difficult") and the number of clients unable to enter (false rejection) will increase.

### 5.4 Experimental Results

Figures 1 and 2 illustrate the performance of SIRP-based speaker identification using feature vectors extracted from concatenations of utterances and averaging of feature vectors from individual utterances. For a large population, SIRP-based identification is approximately three-times more accurate than random guessing but obviously does not yield the accuracy of current methods. These results demonstrate a first step for potential use of SIRPs in speaker identification. Usage might include an adjunct role where SIRPs could supplement existing methods to further improve identification or be used to reduce the parameter requirements of existing methods while maintaining accuracy rates.

A receiver operating characteristic (ROC) curve, initially used in the field of psychophysics, illustrates the performance of a verification methodology over a range of thresholds [6]. The ROC for the present SV experiments is shown in Figure 2. The ROC plots the probability of correct acceptance as a function of the probability of false acceptance. (It should be noted that these two probabilities are not complementary, i.e. $p$(false acceptance) $\neq 1 - p$(correct acceptance), however $p$(false acceptance) $= 1 - p$(correct rejection) and $p$(false rejection) $= 1 - p$(correct acceptance).) Each point on the ROC curve corresponds to a particular threshold. Equal error rate (EER) is used to describe biometric verification system performance. EER is the probability of false

rejection under the conditions when $p(FA) = p(FR)$ The diagonal EER line indicates those points were $p(FA) = p(FR)$. The Random Performance line indicates the points where the probability of false acceptance equals the probability of correct acceptance. The intersection of the ROC curve and the EER line, projected onto the horizontal axis, indicates the EER for the system.

The equal error rate for these experiments was determined to be approximately 38.5%. This can also be seen as the intersection between the ROC curve and the EER line in Figure 2. This contrasts with other methods that report EERs from less than 0.5% to approximately 2% [5]. While a practical SV system would require an EER close to the upper left corner of this plot (low probability of false acceptance and high probability of correct acceptance), the resulting ROC curve does show that the method used provides better than random performance.

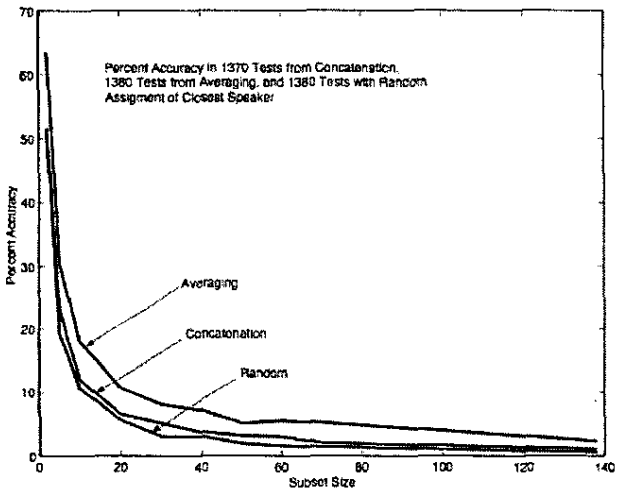


Figure 1: SIRP-based speaker identification performance using both averaging and concatenation methods.

## 6. CONCLUSIONS

As shown in the receiver operating characteristic curve in Figure 2, performance better than random has been demonstrated using 3-element feature vectors. The complexity of computation for these experiments was not excessive. The accuracy for the experiments reported herein does not approach that of other current methods that employ feature vectors comprising up to thousands of elements, however the results suggest that further investigation into the practical feasibility of these small feature vectors may be warranted.

Best performance was obtained using averages of $b_1$, $b_2$, and spectral discriminator over individual utterances, and testing with a simple vector distance method. For each verification session this means averaging over the 4 utterances in each session. For enrollment this means averaging over the 72 utterances comprising the first 3 enrollment sessions for each speaker.

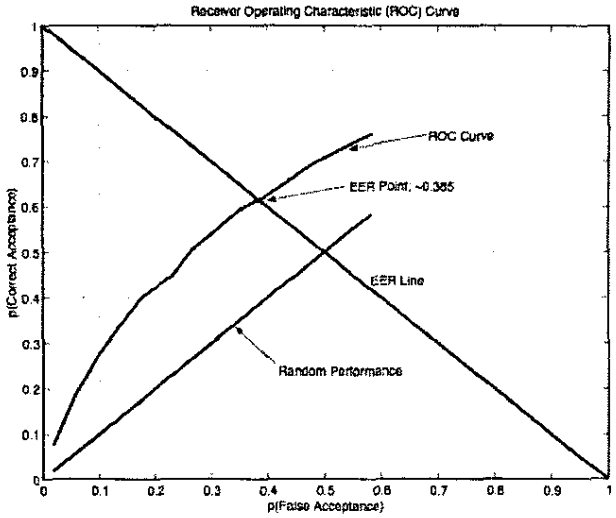


Figure 2: Receiver operating characteristic (ROC).

## 7. REFERENCES


[1] W. Campbell and K. Assaleh, "Polynomial classifier techniques for speaker verification," in *Proc. IEEE ICASSP*, 1999.

[2] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Signal Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] M. Birnbaum, K. Brown, and S. Bardenhagen, "Text-independent speaker identification using fenonic speaker Markov modeling," in *Proc. IEEE ICASSP*, 1996.

[4] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech- model signals," *Signal Processing*, no. 12, pp. 119–141, 1987.

[5] P. De Leon and H. Jiang, "Parameter distributions for speech signals modeled with spherically invariant random processes," in *Proc. 42nd Midwest Symposium on Circuits and Systems*, 1999.

[6] S. Furui and A. Rosenberg, "Speaker verification", in *The Digital Signal Processing Handbook*, CRC Press/IEEE Press, 1998.

[7] J. Campbell, Jr., "Testing with the YOHO CD-ROM voice verification corpus", U.S. Department of Defense, R2, Fort Meade, Maryland.