Parameter Distributions for Speech Signals Modeled with Spherically-Invariant Random Processes

Phillip De Leon and Hui Jiang
New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, NM 88003
Email: {pdeleon, hjiang}@nmsu.edu

***Abstract*— Spherically Invariant Random Processes (SIRPs) have been used as a statistical model for speech and shown to better fit measured probability density functions. With the G-function, SIRPs are conveniently represented and applications involving speech (SIRPs) can be analyzed. This paper describes work in characterizing G-function SIRP parameter distributions for American English.**

I. Introduction

In many DSP applications that process speech signals, it is desirable to have a statistical model for these signals in order to predict system performance. Examples of such applications are speech coders, adaptive filters, and co-channel speech separators [3], [5]. One important aspect of the statistical model for speech is the probability density function (pdf) which mathematically describes the histogram of amplitudes for a large number of speech samples. Standard statistical models for speech pdfs include Laplacian, K_0 (Bessel), and Gamma pdfs as illustrated in Fig. 1 [8]. None

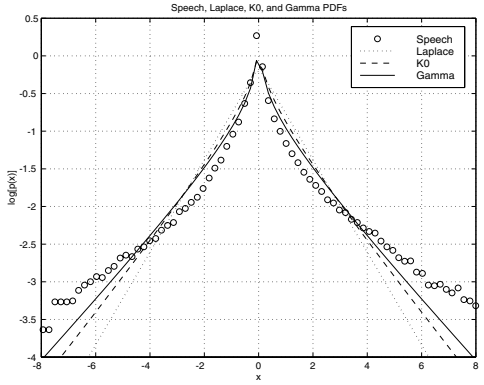

Fig. 1. *Speech pdf and Laplace, K_0, and Gamma pdfs.*

of these pdfs precisely model the speech pdf or the corresponding bivariate (three dimensional amplitude histogram taken from sample pairs of speech signals spaced less than 5ms apart) speech pdf [4]. The bivariate speech pdf captures important temporal relations between the speech samples.

Work by Brehm and Stammler in the 1980's developed of a more refined statistical model for speech signals using G-function Spherically Invariant Random Processes (SIRPs) also known as circularly or spherically symmetric random processes [1], [2], [7]. These SIRPs were shown to model the univariate pdf more precisely than other models and additionally, modeled the bivariate speech pdf as well. The use of SIRPs to model speech signals is based on two important facts: 1) many random processes are SIRPs including those with Laplace and Gamma pdfs and 2) actual speech bivariate pdfs have been shown to exhibit circular or elliptical contours which is a property of the SIRP [2]. G-functions are used to efficiently describe, analyze, and perform calculations on the SIRPs. When used to model speech signals, the G-function description of the SIRP uses two parameters to adjust the pdf to match observed histograms for individual speech signals.

In this paper, we provide an overview on SIRPs and G-functions. We then describe results for G-function SIRP parameter distributions for American English speech based on the TIMIT Speech Corpus [6]. With these parameter distributions, more precise analysis of system performance under American English speech signals can be determined. To illustrate the utility of the results, we calculate a robustness measure on a recently-proposed algorithm for co-channel speech separation and compare to a database of actual speech signals [5].

II. Using SIRPs to Model Speech Signals

A. SIRP Basics

A SIRP can be classified as a stationary random process characterized by having a multivariate pdf $f(\mathbf{x})$, where $\mathbf{x} = [x_1 \dots x_\nu]^T$ that depends only on a radius, $r = \mathbf{x}^T\mathbf{x}$. Thus the ν-order pdf will have circular contours. The most familiar SIRP is the uncorrelated Gaussian process which can be seen by considering the bivariate ($\mathbf{x} = [x_1 x_2]^T$) case (zero mean and unit variance)

$$f(\mathbf{x}) = \frac{1}{2\pi}\exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right). \quad (1)$$

When correlated processes are considered, the SIRP definition is modified to include the correlation matrix, $\mathbf{R}$ and the radius is now defined by $r = \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}$. In the correlated case, the ν-order pdf will have elliptic contours. As an example of a correlated SIRP, consider the bivariate correlated Gaussian process (zero mean, unit variance)

$$f(\mathbf{x}) = \frac{1}{2\pi|\mathbf{R}|^{1/2}} \exp\left(-\frac{\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}}{2}\right). \quad (2)$$

SIRPs are thus completely characterized by their univariate pdf and correlation matrix. In order to analyze systems under SIRPs and still maintain mathematical tractability, the G-function is used to express the univariate SIRP. The next section gives a brief review of the G-function.

B. Expressing SIRPs with the G-Function

The formal definition of the G-function is

$$G\begin{smallmatrix} m & n \\ p & q \end{smallmatrix}\left(z \middle| \begin{smallmatrix} a_p \\ b_q \end{smallmatrix}\right) = \int_C \frac{\prod_{i=1}^{m}\Gamma(b_i - s)\prod_{i=1}^{n}\Gamma(1 - a_i + s)}{\prod_{i=m+1}^{q}\Gamma(1 - b_i + s)\prod_{i=n+1}^{p}\Gamma(a_i - s)} z^s ds \quad (3)$$

where $\Gamma()$ is the well-known Gamma function and the path of integration, C, goes from $\sigma - j\infty$ to $\sigma + j\infty$ so that all poles of $\Gamma(b_i - s)$, $i = 1, \ldots, m$, lie to the right of the path, whereas all poles of $\Gamma(1 - a_i + s)$, $i = 1, \ldots, n$ lie to the left [1]. Fortunately, many operations on the G-function can be reduced to several rules and thus it is not always necessary to work directly with (3) [1].

C. Modeling Bandlimited Speech with SIRPs

The motivation for modeling speech with SIRPs (expressed with G-functions) is for a more precise fit to the univariate speech pdf as well as a fit to the bivariate speech pdf. As developed in [2], the univariate speech pdf is modeled by

$$f(x) = AG\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\left(\lambda x^2 \middle| b_1, b_2\right) \quad (4)$$

where

$$A = \frac{\lambda^{1/2}}{\Gamma(\frac{1}{2} + b_1)\Gamma(\frac{1}{2} + b_2)}$$
$$\lambda = (\frac{1}{2} + b_1)(\frac{1}{2} + b_2). \quad (5)$$

TABLE I
Various pdfs expressed with G functions

pdf	$f(x)$	b_1	b_2	A	λ
Laplace	$\frac{1}{\sqrt{2}}\exp\left(-\sqrt{2}\,\lvert x\rvert\right)$	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2\pi}}$	$\frac{1}{2}$
K_0	$\frac{1}{\pi}K_0(\lvert x\rvert)$	0	0	$\frac{1}{2\pi}$	$\frac{1}{4}$
Gamma	$\left(\frac{\sqrt{3}}{8\pi\lvert x\rvert}\right)^{1/2}\exp\left(-\frac{\sqrt{2}\lvert x\rvert}{2}\right)$	$-\frac{1}{4}$	$\frac{1}{4}$	$\frac{\sqrt{3}}{4\pi\sqrt{2}}$	$\frac{3}{16}$

The $G\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\left(\lambda x^2 \middle| b_1, b_2\right)$ class of functions is a generalization of SIRPs and includes the Gaussian, Laplace, K_0, and Gamma pdfs as described in Table I.

Since the $G\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\left(\lambda x^2 \middle| b_1, b_2\right)$ function provides a general representation for pdfs (some of which have been used to model speech pdfs) there is the possibility that by continuous variation of b_1 and b_2, a whole family of modeling pdfs may be generated—one of which may be a better approximation to the speech pdf than those listed in Table I. This idea is illustrated in Fig. 2 where we have selected b_1 and b_2 to best fit (in a least squared error sense) the speech pdf. We note that in terms of total squared-error between the actual speech pdf and the model, the G-function SIRP model is the best fit. Previous work provided b_1, b_2 parameter data based on five speakers [2]. The data is of limited value due to the small population study and in addition, no distributions for these parameter pairs are given. We present more comprehensive parameter distributions based on a larger population set in the next section.

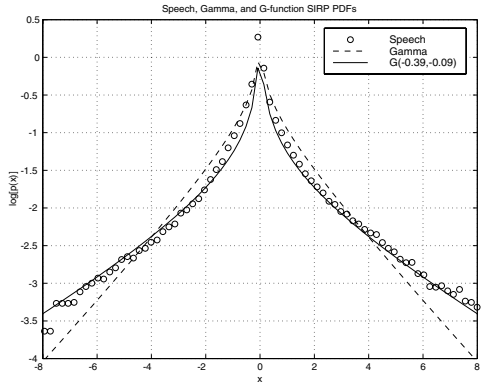

Fig. 2. *Speech pdf and Gamma and G-function SIRP pdf models.*

The bivariate SIRP pdf for modeling the bivariate speech pdf can be computed as

$$f(\mathbf{x}) = \left(\frac{\lambda}{\pi}\right)^{1/2}\frac{A}{|\mathbf{R}|^{1/2}}\times$$

$$G_{1}^{3}{}_{3}^{0}\left(\lambda \mathbf{x}^T \mathbf{R}^{-1} \mathbf{x} \,\middle|\, \begin{matrix} -\frac{1}{2} \\ 0, b_1 - \frac{1}{2}, b_2 - \frac{1}{2} \end{matrix}\right) \quad (6)$$

where the correlation matrix is given by

$$\mathbf{R} = \begin{bmatrix} r_0 & r_N \\ r_N & r_0 \end{bmatrix} \quad (7)$$

and N is the time shift (in samples) between x_1 and x_2 [2].

When the time shift between the two speech signals used to compute the bivariate speech pdf is less than 3.8 ms, the SIRP model also fits the observed contour lines in the bivariate speech pdf very well (Fig. 3) [2]. We note that without extremely long speech signals, it is difficult to acquire enough data points to produce continuous contours of the bivariate speech pdf.

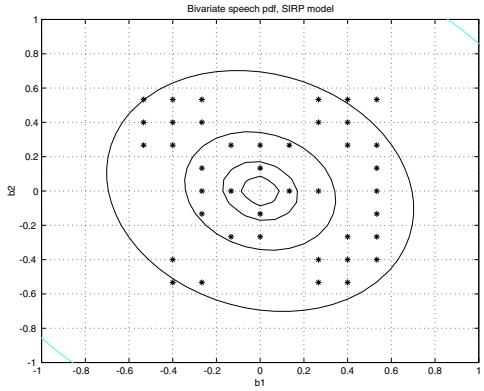

Fig. 3. *Speech bivariate pdf (speech signals 1.25ms apart) (*) and G-function SIRP model (solid line).*

III. SIRP Parameter Distributions for American English

Using the TIMIT speech corpus of 630 speakers, we computed G-function SIRP b_1, b_2 parameters for each speech signal. The parameters are determined by first computing the histogram of each speech signal. Then a search for the b_1, b_2 parameter pair which best fits the histogram is conducted. We have found in our work that the error surface from such a search has a single global minimum so a unique parameter pair is obtainable.

With the b_1, b_2 parameter pairs from each of 630 speakers collected, we are able to characterize the parameter plane for the G-function SIRP model of American-English. Figs. 4 and 5 illustrate this characterization. We note from the data, that various proportions of the TIMIT population are well represented within a small parameter range as described in Table II.

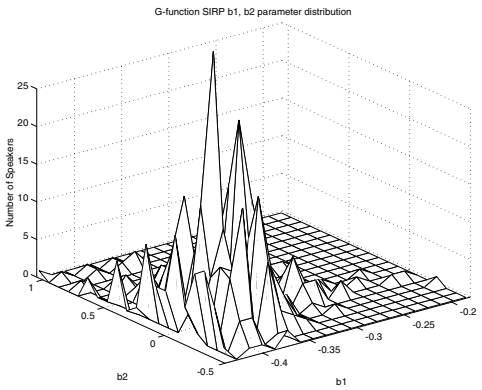

Fig. 4. *Distribution of b_1, b_2 parameter pairs for TIMIT speech corpus.*

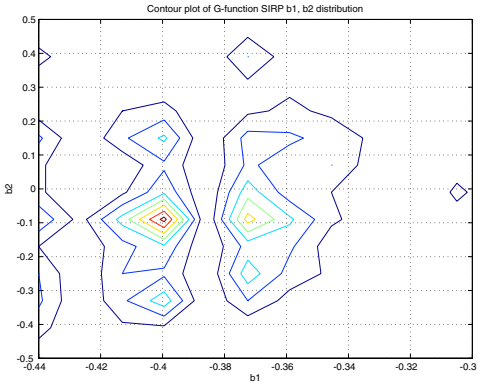

Fig. 5. *Contour plot of distribution of b_1, b_2 parameter pairs for TIMIT speech corpus.*

IV. Robustness Calculation for Speech Separation Algorithm

The problem of co-channel (mixed) speech separation has been recently studied and one proposed method for separation employs a strategy of *kurtosis maximization* [5]. This work relies on a critical assumption that the kurtosis, defined as

$$\kappa_X = \frac{E\left[x^4\right]}{E\left[x^2\right]^2} \quad (8)$$

is lower in the mixture of speech signals than in the separated speech signals.

TABLE II
b_1, b_2 ranges for TIMIT population.

100% of TIMIT population	$-0.44 \leq b_1 \leq -0.17$ $-0.49 \leq b_2 \leq 1.11$
95% of TIMIT population	$-0.42 \leq b_1 \leq -0.26$ $-0.49 \leq b_2 \leq 0.5$
90% of TIMIT population	$-0.41 \leq b_1 \leq -0.27$ $-0.49 \leq b_2 \leq 0.42$
75% of TIMIT population	$-0.41 \leq b_1 \leq -0.28$ $-0.29 \leq b_2 \leq 0.32$

Since we assume zero-mean, unit-variance SIRPs, we need only compute $E\left[x^4\right]$ in (8) since $E\left[x^2\right]^2 = 1$. It can be shown that for the $G\begin{smallmatrix}2 & 0\\ 0 & 2\end{smallmatrix}$ SIRP,

$$\kappa_X = \frac{\left(b_1+\frac{3}{2}\right)\left(b_2+\frac{3}{2}\right)}{\left(b_1+\frac{1}{2}\right)\left(b_2+\frac{1}{2}\right)}. \quad (9)$$

As a side note, using b_1, b_2 values in Table I in (9) we get the well-known kurtosis values for Laplace ($\kappa_X = 6$), K_0 ($\kappa_X = 9$), and Gamma ($\kappa_X = 11\frac{2}{3}$) pdfs.

For a mixture of independent SIRP random variables, $Z = \alpha X + (1-\alpha)Y$ where α is the mixing parameter, it can be shown (assuming zero-mean, unit-variance random processes)

$$\kappa_Z = \left\{ \alpha^4 \frac{\left(b_{x,1}+\frac{3}{2}\right)\left(b_{x,2}+\frac{3}{2}\right)}{\left(b_{x,1}+\frac{1}{2}\right)\left(b_{x,2}+\frac{1}{2}\right)} + 6\alpha^2(1-\alpha)^2 + (1-\alpha)^4 \frac{\left(b_{y,1}+\frac{3}{2}\right)\left(b_{y,2}+\frac{3}{2}\right)}{\left(b_{y,1}+\frac{1}{2}\right)\left(b_{y,2}+\frac{1}{2}\right)} \right\} \Big/ \left[\alpha^4 + 2\alpha^2(1-\alpha)^2 + (1-\alpha)^4\right] \quad (10)$$

where $b_{x,1}$,$b_{x,2}$; $b_{y,1}$, $b_{y,2}$ are G-function parameters associated with SIRP random variables X; Y respectively.

Based on b_1, b_2 data for the TIMIT corpus presented in Section III. we compute the probability that the kurtosis of a mixture of speech signals (modeled with G-function SIRPs) is less than the kurtosis of both individual speech signals. Computing this probability will give us a robustness measure regarding the kurtosis-based speech separation algorithm in which the critical assumption must be satisfied. Mathematically, we must estimate the probability, $P\{\kappa_Z < \min(\kappa_X, \kappa_Y)\}$. In order to estimate this probability, for each pair of TIMIT speakers and associated b_1, b_2 SIRP parameters, we compute (9) for each individual speaker and (10) for the mixture with fixed α. We count the number of pairs which satisfy the critical assumption and divide by the total number pairs. We repeat this procedure for a variety of mixing ratios. The results are illustrated in Fig. 6. In addition, we repeat the same robustness measure using 100 actual speech signals from the TIMIT corpus (Fig. 6). We note from Fig. 6, that the robustness measure using the SIRP model provides a rough lower bound to the robustness measure using actual speech signals. The measures indicate the kurtosis-based separation algorithm will be mostly successful in separating out both speech signals when the mixing ratio is approximately equal. At extreme mixing ratios the algorithm will not be as successful, however, we note that at the extreme mixing ratios, one signal already dominates and is therefore almost naturally separated.

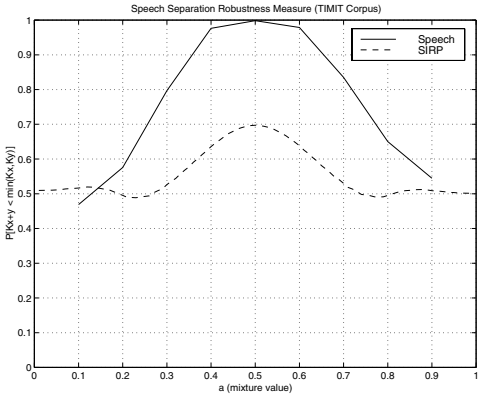

Fig. 6. *Probability of satisfying critical assumption in speech separation algorithm for various mixing ratios using actual speech signals and SIRP model.*

V. Conclusion

In this paper we have reviewed the use of spherically invariant random processes to model speech signals and the G-function description of the model. Using the TIMIT speech corpus of 630 American-English speakers, we have characterized the G-function parameter plane. As an example illustrating the use of this characterization, a robustness measure associated with a recent co-channel speech separation algorithm was computed.

References

[1] H. Brehm, "Description of Spherically Invariant Random Processes by Means of G-Functions," *Lecture Notes in Mathematics*, Springer Verlag, vol. 969, pp. 39–73, 1982.

[2] H. Brehm and W. Stammler "Description and Generation of Spherically Invariant Speech-Model Signals," *Signal Processing*, vol. 12, no. 2, pp. 119–141, Mar. 1987.

[3] M. Rupp "The Behavior of LMS and NLMS Algorithms in the Presence of Spherically Invariant Processes," *IEEE Transactions on Signal Processing*, vol. 41, no. 3, pp. 1149–1160, Mar. 1993.

[4] D. Wolf and H. Brehm "Experimental Studies on One- and Two-Dimensional Amplitude Probability Densities of Speech Signals," *Proceedings of the 1973 International Symposium on Information Theory*, pp. B4–B6.

[5] J. LeBlanc and P. De Leon " Speech Separation by Kurtosis Maximization," *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[6] W. Fisher, G. Doddington, and K. Goudie-Marshall, " The DARPA Speech Recognition Research Database: Specifications and Status," *Proceedings of the DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.

[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill: New York, N.Y., 3ed, pp. 109–112, 1989.

[8] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, PPrentice-Hall: Englewood Cliffs, N.J., pp. 174–179, 1978.