Speech Probability Distribution

Saeed Gazor, *Senior Member, IEEE,* and Wei Zhang

***Abstract*—It is demonstrated that the distribution of speech samples is well described by Laplacian distribution (LD). The widely known speech distributions, i.e., LD, Gaussian distribution (GD), generalized GD, and gamma distribution, are tested as four hypotheses, and it is proved that speech samples during voice activity intervals are Laplacian random variables. A decorrelation transformation is then applied to speech samples to approximate their multivariate distribution. To do this, speech is decomposed using an adaptive Karhunen–Loève transform or a discrete cosine transform. Then, the distributions of speech components in decorrelated domains are investigated. Experimental evaluations prove that the statistics of speech signals are like a multivariate LD. In brief, all marginal distributions of speech are accurately described by LD in decorrelated domains. While the energies of speech components are time-varying, their distribution shape remains Laplacian.**

***Index Terms*—Speech coding, speech processing.**

I. Introduction

MODELING of speech signals is motivated by many applications of speech processing, such as speech coding [3], speech recognition [11], speech enhancement [9], and independent component analysis and computational auditory scene analysis (e.g., see [6] and [12]). In order to design better speech processing systems, it is important to have a reasonable approximation for the probability density function (pdf) of speech and noise. Various statistics of speech signals such as speech kurtosis have been studied in many scenarios and have found a wide range of applications.

In the early 1950s, the pdf of speech signals in the time domain was investigated by Davenport [1]. Early results show that a good approximation is the gamma distribution (γ-D), while a poorer and simpler approximation is reported to be the Laplacian distribution (LD) [2]–[4]. In many applications, the speech signal is assumed to be Gaussian in order to simplify the derivation of speech processing algorithms. More recently, it has been shown that samples of a bandlimited speech signal can be represented by a multivariate Gaussian distribution (GD) with a slowly time-varying power if samples are taken within short time intervals of less than 5 ms [4], [5]. The power variation makes the long-term distribution a γ-D or LD.

As each coefficient of a discrete-time Fourier transform (DFT) is a linear weighted sum of speech samples as random variables, a GD is justified by the central limit theorem for speech signals in the DFT domain. This assumption has been used in several applications, e.g., [10]. As the frame length increases, the pdf of coefficients is believed to converge to GD. In practice, the GD assumption for speech is not an accurate approximation, not only because of the length of data but also because the necessary conditions for the central limit theorem (even in its asymptotical weak form) are not satisfied.

Manuscript received November 18, 2001; revised September 30, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Hasegawa-Johnson.

The authors are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, K7L 3N6 Canada.

Digital Object Identifier 10.1109/LSP.2003.813679

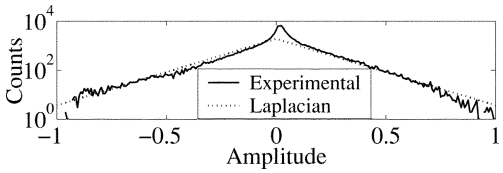


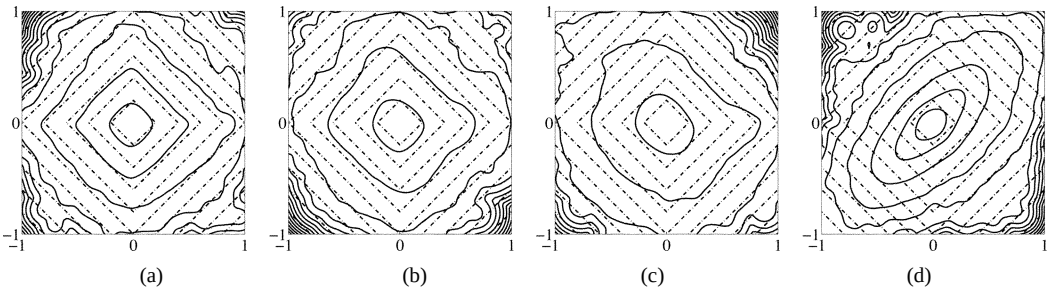
Fig. 1. PDF estimate of a typical speech signal.

The underlying assumption in most speech processing schemes is that the properties of the speech signal change relatively slowly with time. This assumption leads to a variety of processing methods involving intervals of about 1–60 ms, in which frames of the speech signal are isolated and processed as if they had the same fixed properties [9]. The aim of this letter is to investigate the distribution of speech samples in these kinds of time intervals, analogous to investigation of natural images in the DCT domain [7], [8].

II. Time Domain Modeling

In this section, samples of a speech signal are considered as samples of a random variable, and the speech signal's pdf in the time domain is investigated. This is in contrast to the following sections, where the signal is first transformed into the Karhunen–Loève transform (KLT) and DCT domains Subsequently, the joint distribution of the transformed components is investigated. The modeling in the time domain is simple to understand and to use, provides a useful basis for further investigations, and motivates us to model the speech signal in other domains.

Fig. 1 gives the amplitude distribution of a typical speech signal (speech samples are taken from *http://www.dailywav.com* with an SNR >15 dB and about 50% of silence duration) where the y axis is depicted using a logarithmic scale, and the Laplacian curve is plotted after a curve fitting. From the fact that the experimental curve has a linear tail and is symmetrical, it could be concluded that the pdf of this signal follows an LD as indicated by Davenport [1] and Richards [2]. The portions of the curve with least probability cannot be estimated accurately and produce noisy "tails" in the distribution graph of Fig. 1. The difference between the experimental distribution and LD in the central region corresponds to the silence intervals in the speech signal. A generalized Gamma pdf $f_{\mathbf{x}}(x) = |x|^h e^{-|x|/a}/2h!a^{h+1}$ (with $a>0$ and $h>-1$), which includes the LD (i.e., $h=0$) and for $h=-0.5$, commonly referred to as Gamma density, can fit better in the central part of the curve for



(a) (b) (c) (d)

Fig. 2. Comparison of contours for bivariate pdf of two samples of speech $(\mathbf{x}(t), \mathbf{x}(t+\tau))$. (Solid line) Experimental results estimated by $\hat{f}_{\mathbf{x}(t),\mathbf{x}(t+\tau)}(x, y) = 1/N \sum_{t=1}^{N} (1/4\pi\sigma^2) \exp(-(|x - x(t)|^2 + |y - x(t+\tau)|^2)/\sigma^2)$ where $x(t)$ is a long speech signal, and σ^2 is a small number. (Dashed-dotted) LD $f(x, y) = (1/4a^2) \exp(-(|x| + |y|)/a)$. (a) $\tau = 25$ ms. (b) $\tau = 2.5$ ms. (c) $\tau = 1$ ms. (d) $\tau = 0.1$ ms.

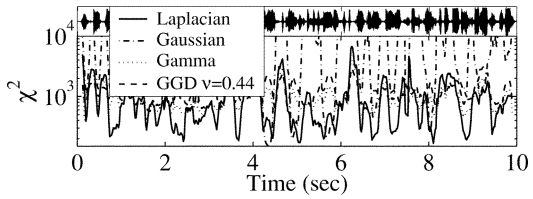


Fig. 3. χ^2 test result over a time interval (200 ms) for four pdf hypotheses.

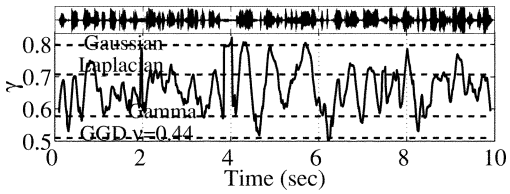


Fig. 4. Moment test ($\gamma = E[|\mathbf{x}|]/\sqrt{E[\mathbf{x}^2]}$) value for a speech signal and the nominal values for different hypotheses over time intervals of 200 ms.

suggested values of h in the range of -0.8 to $+0.5$ [2]. Using a voice activity detector, we omitted silence intervals from the data and noted that the central part fits the LD well. Previously published results [1], [2] show that speech has a γ-D. We conclude that samples of speech signals during voice activity intervals are Laplacian, and speech as a mixture of silences and voice activity intervals (depending on the SNR and the silence duration percentage: $0 \leq p \leq 1$), results in a γ-D.

Fig. 2 illustrates the contours of bivariate pdf of two samples of speech. For time intervals less than 1 ms, it can be seen that the experimental contours are elliptical; thus, the multivariate density of speech is spherically invariant [4], [5]. For time intervals over 2 ms, the contours are more like squares; therefore, we conclude that the bivariate pdf looks like an LD and is not spherically invariant.

One may ask whether only long data samples of speech exhibit an LD. For time intervals greater than 5 ms, using tests described in the Appendixes, our experiments show that among the following widely known speech pdfs [1]–[6] the LD is favored:

- Gaussian pdf: $f_{\mathbf{x}}(x) = 1/(\sqrt{2\pi\sigma^2}) \exp(-(x^2/2\sigma^2))$;
- Laplacian pdf: $f_{\mathbf{x}}(x) = (1/2a) \exp(-(|x|/a))$;
- Gamma pdf: $h = -0.5$, $f_{\mathbf{x}}(x) = |x|^h e^{-|x|/a}/2h!\, a^{h+1}$;
- GGD pdf: $\nu = 0.44$ (see Appendix B).

First, the χ^2 test value (see Appendix A) is calculated for the above distributions for each 200–ms time frame (for shorter frames, results are similar) where the overlap between successive frames is 100 ms. The χ^2 test result for a speech signal is displayed in Fig. 3. The lower the χ^2 value the better the corresponding hypothesis fits the data. Thus, from Fig. 3, we conclude that the best fit is obtained under the hypothesis that the speech has an LD. We note that only during silent intervals might the χ^2 value for the Gaussian hypothesis be slightly less than that of the Laplacian hypothesis; that is again to say that speech has an LD during voice activity time intervals.

Both GD and LD functions are special cases of the generalized Gaussian distribution (GGD) function [6], [8]. In Appendix B, a moment test is developed to evaluate these hypotheses. For each speech frame as used in the χ^2 test, we calculated the sample variance $E[\mathbf{x}^2]$ and the sample mean of the absolute value $E[|\mathbf{x}|]$. The test statistic $\gamma = E[|\mathbf{x}|]/\sqrt{E[\mathbf{x}^2]}$ is then calculated and depicted in Fig. 4. Each distribution has a different nominal test statistic for γ; these are illustrated by four straight lines. We can see that the LD is a better approximation than the others. Again, Fig. 4 shows that the shape of the distribution (γ) is closer to LD during active speech intervals compared to silent intervals. The γ-D is the next candidate in the frames that are a mixture of silence and voice activity. The moment test favors GD for silent intervals.

TABLE I
AVERAGE χ^2 VALUE FOR DIFFERENT WINDOW LENGTHS

Frame Length	GGD, $\nu = 0.44$	γ–D	LD	GD
2.5ms	154	113	78	[68]
5ms	194	140	[96]	124
10ms	272	192	[127]	245
20ms	407	284	[181]	562
50ms	688	496	[350]	4502
100ms	912	725	[717]	26477
200ms	1151	1013	[926]	31477
0.5s	1857	1739	[1726]	72665
1s	3093	[2910]	2971	50886
2s	5717	[5525]	5773	46000
5s	[14725]	14750	15692	80330

Table I illustrates the χ^2 test value of a 20-s speech signal for different pdf hypotheses and frame lengths. The χ^2 value is calculated within each frame and is averaged over time. Table I shows that the speech signal has an LD within time frames longer than 5 ms and a GD for time frames of less than 2.5 ms. For time frames longer than 0.5 s, a γ-D or GGD with $\nu = 0.44$ is favored, because long frames usually consist of both voiced and silent samples. This behavior is justified by Fig. 5, which depicts variations of two moment values of an LD-GD mixture.

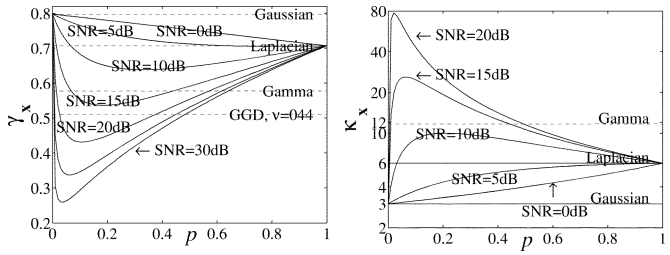


Fig. 5. Moment test values $\gamma_{\mathbf{x}} = E[|\mathbf{x}|]/\sqrt{E[\mathbf{x}^2]}$ and $\kappa_{\mathbf{x}} = E[\mathbf{x}^4]/E[\mathbf{x}^2]^2$ of a random selection of $\mathbf{v} \sim$ LD (with probability p) or $\mathbf{u} \sim$ GD (with probability $1-p$), i.e., $\mathbf{x} \sim (1-p)f_{\mathbf{u}}(x) + pf_{\mathbf{v}}(x)$, SNR $= E[\mathbf{v}^2]/E[\mathbf{u}^2]$.

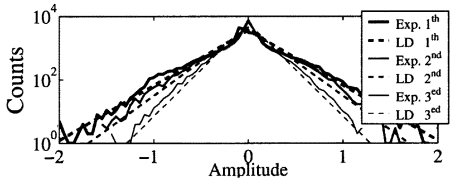


Fig. 6. PDF of three dominant speech signal components in the adaptive KLT domain proposed in [9]. (Solid line) Experimental. (Dashed line) LD.

These moments, depending on the SNR and the silence duration percentage $0 \leq p \leq 1$, are similar to those of γ-D, GGD with $\nu = 0.44$, GD, or LD. This is why in some of the literature γ-D is believed to be the best model, while in other references GGDs with $\nu = 0.44$ or LD are favored [1]–[6].

III. KLT Domain Modeling

We have seen that time domain samples of a speech signal are very well described by an LD. This gives a first-order pdf characterization for speech. In order to obtain a higher order statistical description, we first transform the speech signal into uncorrelated components. We then show that these components themselves are distributed like Laplacian random variables with different parameters. The speech signal is regarded as a nonstationary process, and yet we will see that the shapes of the pdf of its components are almost time-invariant and are LDs with slowly time-varying parameters.

A KLT is used here to decorrelate the data vector and to separate the speech components [9], [11], [12]. It is assumed that successive speech samples could be accurately modeled by the pdf of the KLT components that are uncorrelated random variables. This modeling is described by the KLT and the energy (or mean absolute value) of the KLT components. Speech is often viewed as a nonstationary process; therefore, this transformation varies slowly with time and should be updated with the arrival of new samples. In this letter, the KLT is adaptively estimated by the algorithm proposed in [9]. We consider the three most dominant components of a speech signal in the KLT domain obtained by the algorithm in [9]. Fig. 6 shows that the pdfs of these components (that are uncorrelated) for a 10-s speech sample are approximated very well by Laplacian pdfs with different parameters. Note that minor components that have very small energies are neglected in practice.

To further investigate the behavior of speech components, the χ^2 test is similarly applied to these three components. The results are depicted in Fig. 7, where the χ^2 parameter for each KLT component is also computed over every 200-ms time frame with 100-ms overlaps between frames. Again, we see that the KLT

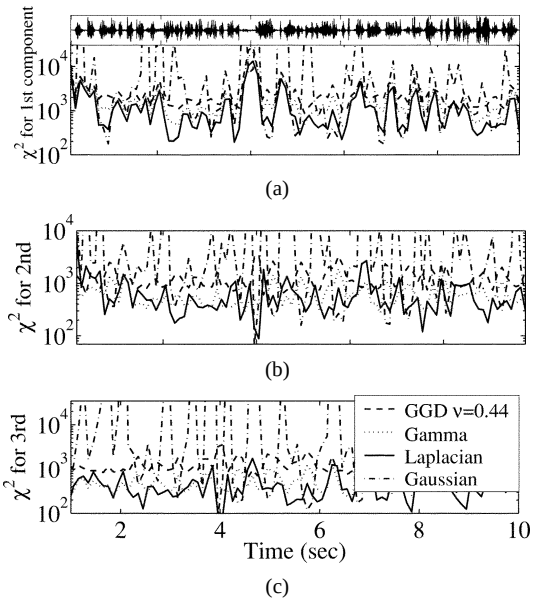


Fig. 7. Short-time (200 ms) χ^2 test results for KLT coefficients.

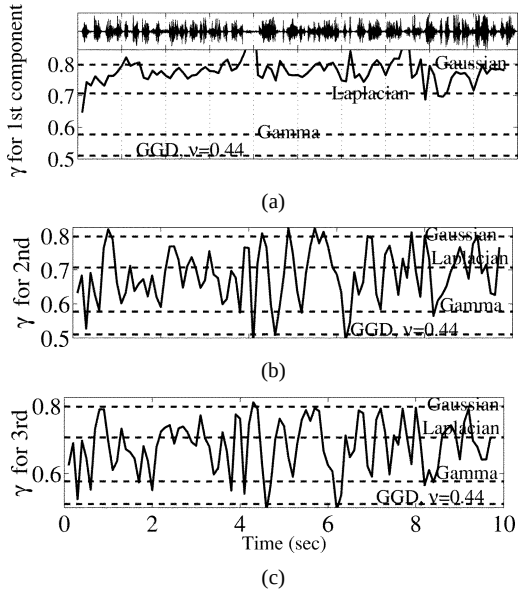


Fig. 8. Moment test results for DCT coefficients of a speech signal.

coefficients are better approximated by the LD rather than by the others. Hence, the joint pdf of samples of a speech signal could be approximated by assuming that speech in the KLT domain is a multivariate random vector with uncorrelated LD elements and different, slowly time-varying parameters.

IV. DCT Domain Modeling

The KLT is time varying, data dependent, and needs to be updated for each new speech sample; therefore, simple orthogonal transformations like FFT are used instead in many applications. Of all discrete orthogonal transforms, the DCT is the most popular, not only for its nearly optimal performance in whitening lowpass signals, but also for its computational efficiency.

Speech signal vectors are transformed using DCT. Then, the moment test is applied to each DCT coefficient. The results for three components are depicted in Fig. 8. Fig. 8(a) shows that

only the first coefficient (i.e., the discrete cosine (DC) component) is better approximated by the GD hypothesis. We must note that this very small low-frequency component is not representative of a speech signal, because the DC component is usually blocked in signal acquisition and represents just noise measurement. We conclude from Fig. 8(b) and (c) (as well as from all other components) that the DCT coefficients of speech are better approximated by LD. This is to say that speech in the DCT domain also exhibits an LD, with the result that a multivariate LD for speech is more accurate than a Gaussian one.

V. Discussion and Conclusion

In previously published results, the distribution of the speech signal is claimed to have a γ-D over long time periods [1]–[3], while in some other papers a GD is used. The speech is a spherically invariant random process for time intervals less than 2 ms [4]. We first demonstrate that a speech signal *during voice activity intervals* may be characterized by LD in the time domain. Then two different tests are used to compare widely known speech distributions over short time periods. Both tests favor LD *during voice activity intervals*. Tests performed in the KLT and DCT domains lead us to conclude that all major and meaningful components of voiced speech show LD.

In decorrelated transformed domains such as the KLT and DCT, speech components are uncorrelated. Our experimental results suggest a multivariate Laplacian pdf for speech during its activity intervals. This pdf is simple to use and fits better than other widely used pdfs in many speech processing applications.

In contrast, similar tests performed on noise gathered from different acoustic environments clearly illustrate that most of them are better described by a GD. During silent intervals, both criteria (i.e., γ and χ^2) favor the GD hypothesis.

We also noted that these tests, for those frames containing both speech and silence (on the edges), favor γ-D similar to the expected results from a random mixture selection of a GD and an LD. This shows that the difference between our results and previously published results [1]–[6] is caused by the inclusion in previous studies of silent samples (i.e., low-power noise, which is well described by a GD) mixed with active speech (which is described by an LD) (see Fig. 5).

Appendix A
χ^2 Test

The χ^2 test compares experimental data with some given pdfs and measures the distortion between the pdfs and data by $\chi^2 = \sum_{i=1}^{K} (n_i - Np_i)^2/Np_i$ where the sample space is partitioned into K intervals; n_i is the number of samples that fall into the ith interval; p_i is the theoretical probability that a sample will fall into the same ith interval; and N is the total number of samples. The smaller the χ^2 value, the better the fit.

Appendix B
Moment Test for GGD and γ-D

Assume that the pdf of the speech signal $\mathbf{x}$ is modeled with the zero-mean GGD function

$$f_{\mathbf{x}}(x) = \frac{\nu\alpha(\nu)}{2\sigma\Gamma(1/\nu)} \exp\left\{-\left[\alpha(\nu)\frac{|x|}{\sigma}\right]^{\nu}\right\}$$

where $\alpha(v) = \sqrt{(\Gamma(3/\nu)/\Gamma(1/\nu))}$ and $\Gamma(\bullet)$ denotes the Gamma function, and ν and σ are positive real parameters. The shape parameter ν describes the exponential rate of decay; σ is the square root of the variance. If $\nu = 1$, the above is an LD, and if $\nu = 2$, it is a GD. The maximum-likelihood estimation of the GGD parameters can be found in [8]. The shape parameter of a GGD can be estimated by $\hat{\nu} = F^{-1}(\gamma)$ where $\gamma = E[|\mathbf{x}|]/\sqrt{E[\mathbf{x}^2]}$, $E[\mathbf{x}^2]$, $E[|\mathbf{x}|]$, and $F(\gamma) = \Gamma(2/\gamma)/\sqrt{\Gamma(1/\gamma)\Gamma(3/\gamma)}$ [7]. The calculation of F^{-1} is complex and time consuming in real-time applications such as speech processing. As we can easily show that $F^{-1}(\gamma)$ is a monotonic function increasing for γ for the interval of interest, we suggest using γ instead of ν as a moment value to perform the test.

This test evaluates a sample set and finds its best pdf in the class of GGD functions. Under the hypothesis that the shape parameter ν is close to one, the pdf is Laplacian, and if ν is close to two, the pdf is Gaussian. Thus, using γ instead of ν as a decision parameter, the value of γ is to be compared with $F(1) = 1/\sqrt{2}$ and $F(2) = \sqrt{2/\pi}$, to verify whether the sample set should be considered an LD or GD. In our simulations, the case $\nu = 0.44$ is also considered where $F(0.44) = 0.5104$. The moment value for the Gamma distribution is $1/\sqrt{3}$.

Acknowledgment

Authors would like to thank Associate Editor, the anonymous reviewers, and H. Warder for their comments and suggestions.

References

[1] W. B. Davenport, "An experimental study of speech wave probability distributions," *J. Acoust. Soc. Amer.*, vol. 24, no. 4, pp. 390–399, July 1952.

[2] D. L. Richards, "Statistical properties of speech signals," *Proc. Inst. Elect. Eng.*, vol. 111, no. 5, pp. 941–949, 1964.

[3] M. D. Paez and T. H. Glisson, "Minimum-mean squared-error quantization in speech PCM and DPCM," *IEEE Trans. Commun.*, vol. COM-20, pp. 225–230, Apr. 1972.

[4] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signal," *Signal Process.*, vol. 12, no. 2, pp. 119–141, Mar. 1987.

[5] J. P. LeBlanc and P. L. De Leòn, "Speech separation by kurtosis maximization," in *Proc. ICASSP*, vol. 2, 1998, pp. 1029–1032.

[6] G.-J. Jang, T.-W. Lee, and W.-H. Oh, "Learning statistically efficient features for speaker recognition," in *Proc. ICASSP*, 2001.

[7] R. L. Joshi and T. R. Fischer, "Comparison of generalized Gaussian and Laplacian modeling in DCT image coding," *IEEE Signal Processing Lett.*, vol. 2, pp. 81–82, May 1995.

[8] F. Müller, "Distribution shape of two-dimensional DCT coefficients of natural images," *Electron. Lett.*, vol. 29, no. 22, pp. 1935–1936, Oct. 1993.

[9] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.

[10] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, pp. 1–3, Jan. 1999.

[11] J. Huang and Y. Zhao, "A DCT-based fast signal subspace technique for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 747–751, Nov. 2000.

[12] J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.