A STUDY OF THE DISTRIBUTION OF TIME-DOMAIN SPEECH SAMPLES AND DISCRETE FOURIER COEFFICIENTS

Jesper Jensen, Ivo Batina, Richard C. Hendriks, and Richard Heusdens

{J.Jensen, I.Batina, R.C.Hendriks, R.Heusdens}@ewi.tudelft.nl
Department of Mediamatics
Delft University of Technology
2628 CD Delft, The Netherlands

ABSTRACT

We study the distribution of time-domain speech samples as well as the distribution of Discrete Fourier Transform (DFT) coefficients obtained from speech segments. We consider four possible pdf model types, namely Gaussian, Laplacian, Gamma, and a Generalized Gaussian density (GGD). Our time-domain results suggest that for segment lengths of 20–200 ms, the Laplacian density is the better choice, while for shorter segments the Gaussian model is more appropriate. For segments of 20 ms, the Gaussian model is advantageous for broad speech classes of fricatives, nasals and glides, but stop sounds are much better represented with the Laplacian model, making the latter model better on average. Finally, our study supports the often made assumption that DFT coefficients collected within short time intervals can be considered Gaussian distributed, across all types of speech sounds.

1. INTRODUCTION

Statistical signal processing methods have found widespread use in numerous digital speech applications including speech coding [1], recognition [2] and enhancement, e.g. [3]. Typically, these methods consider the speeech signal as a realization of a random process and rely on assumed or estimated statistical models to describe this process (or functions thereof). Usually, the accuracy of the assumed statistical model affects directly the performance of the algorithm in question. However, the choice of underlying statistical model is not only governed by the desire to have accurate models, but also by the need for simple, mathematically tractable representations.

Assuming independent and identically distributed (iid) time domain speech samples, and focusing on the long-term probability distribution[1], it has been shown that the two-sided Gamma distribution of the form

$$f_{x_n}(x) = \left(\frac{\sqrt{3}}{8\pi\sigma_x|x|}\right)^{1/2} \exp\left(-\frac{\sqrt{3}|x|}{2\sigma_x}\right), \qquad (1)$$

where $-\infty < x < \infty$ and σ_x denotes the standard deviation, provides a good approximation of the underlying pdf [4, 5, 6]. Moreover, the study in [7] suggests that a generalized Gaussian density (GGD)[2] of the form

$$f_{x_n}(x) = \frac{\nu\alpha(\nu)}{2\sigma_x\Gamma(1/\nu)} \exp\left\{-\left(\alpha(\nu)\frac{|x|}{\sigma_x}\right)^{\nu}\right\}, \qquad (2)$$

where $-\infty < x < \infty$, $\Gamma(\cdot)$ denotes the Gamma function, and $\alpha(\nu) = \sqrt{\Gamma(3/\nu)/\Gamma(1/\nu)}$, provides an even better model for the long-term description, if the shape parameter ν is chosen as $\nu = 0.44$. For time domain speech samples collected over shorter segment lengths in the order of 5-500 ms, it was argued in [7] that the somewhat simpler Laplacian (two-sided exponential) density

$$f_{x_n}(x) = \frac{1}{\sqrt{2}\sigma_x} \exp\left(-\frac{\sqrt{2}|x|}{\sigma_x}\right), \quad -\infty < x < \infty \qquad (3)$$

is a better approximation than the Gamma or GGD models. In [6] it was stated that short-term pdf (without specifying exactly the meaning of *short-term*) of speech segments can be approximated well by a Gaussian pdf

$$f_{x_n}(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2\sigma_x^2}x^2\right), \quad -\infty < x < \infty. \qquad (4)$$

A similar conclusion was drawn in [7], where it was shown that for signal segments with a duration shorter than 5 ms, a Gaussian pdf provides a better fit than the Laplacian model.

Since many speech processing algorithms operate in transform domains rather than directly on the time domain samples, the distribution of various transform coefficients has also been studied, see e.g. [8, 9] for some results on the distribution of Discrete Fourier Transform coefficients, and [7] for a study of the distribution of Discrete Cosine Transform (DCT) and Karhunen Loéve Transform (KLT) coefficients. Observing that DFT coefficients are merely (complex-valued) linear combinations of time samples, and assuming the time samples to be Gaussian and/or independent, speech DFT coefficients have often been assumed to obey a complex Gaussian distribution[3], see e.g. [11]. The study in [8, 9], however, suggests that the DFT coefficient distribution is better approximated using a Gamma or Laplacian density.

[1] Long-term refers here to the fact that the probability distributions were estimated from histograms of time domain speech samples collected over tens of seconds of speech material.

[2] We see that the GGD contains as a special case the Gaussian density ($\nu = 1$) and the Laplacian density ($\nu = 2$).

[3] It is easy to verify that a linear combination of Gaussian random variables is also Gaussian, while the Central Limit Theorem [10] ensures that a linear combination of independent variables approaches a Gaussian density.

In this work we study the distribution of time domain speech samples as well as DFT coefficients. Our goal is not only to verify the findings of the work refered above, but also to provide a more detailed analysis of the distributions with respect to analysis segment lengths, speech sounds, and speaker gender.

2. ANALYSIS OF PROBABILITY DISTRIBUTIONS

We consider four possible candidate distributions

- c_1 Gaussian, Eq. (4).
- c_2 Laplacian, Eq. (3).
- c_3 Gamma, Eq. (1).
- c_4 GGD ($\nu = 0.44$), Eq. (2).

We focus on these distributions because they appear to cover the field of assumed pdf models most often encountered in literature, e.g. [11, 8, 7, 9, 12]. In the following we review and derive methods for determining the goodness-of-fit between an observed data sequence and a hypothesized probability density function.

2.1. Chi-Square Goodness-of-Fit Test

Let $x = [x_0 \dots x_{N-1}]$ denote a sequence of observed data samples, which we assume are iid, and let $f(x)$ denote some postulated or assumed model pdf $f(x) = \prod_{n=0}^{N-1} f(x_n)$, where $f(x_n)$ denotes the pdf of any sample in x. The chi-square test is one of several widely used methods for determining the goodness-of-fit of the postulated pdf to the observed data, see e.g. [13]. Consider the hypotheses:

- H_0: The observed sequence x is distributed according to a specified distribution $f(x)$.
- H_1: The observed sequence does not follow $f(x)$.

Our goal is to test if hypothesis H_0 can be rejected at a given significance level α, where α defines the probability of rejecting H_0 when it is true. Let X denote the input space of any element x_n in x, and partition this input space into a union of K disjoint regions X_k such that

$$X = \cup_{k=1}^{K} X_k.$$

Let N_k denote the number of samples of x that fall in region X_k, and denote by E_k the expected number of observations in region X_k under the assumption that the samples x_n are drawn from the postulated pdf, i.e., $E_k = N \int_{X_k} x_n \, f(x_n) dx_n$. Define the test statistic D^2 as follows:

$$D^2 = \sum_{k=1}^{K} \frac{(N_k - E_k)^2}{E_k}. \qquad (5)$$

Clearly, $D^2 \geq 0$ is a measure of the fit between the observed and the postulated pdf; lower D^2 values imply a better fit, while large D^2 values suggest that H_0 should be rejected. More formally, we reject H_0 if $D^2 \geq t_{\alpha,\chi^2}$, where t_{α,χ^2} is a threshold determined by the significance level of the test. In order to find the threshold value t_{α,χ^2}, we use the observation that for large N, the distribution of the random variable D^2 approaches a chi-square distribution with $K-1$ degrees of freedom [13]. Thus, the threshold t_{α,χ^2} can be computed by finding the point at which $\int_{t_{\alpha,\chi^2}}^{\infty} \chi^2_{K-1}(y)dy = \alpha$, where $\chi^2_{K-1}(y)$ denotes a (scalar) chi-square pdf with $K-1$ degrees of freedom.

The discussion so far has assumed that the postulated distributions were completely specified. However, in the problem at hand, we are given an observed sequence of data samples x and a postulated pdf *type*, e.g. zero-mean, Gaussian. Thus, in order to completely specify the pdf, we determine the pdf model parameters from the data using maximum likelihood estimates. In doing so, it has been observed that the distribution of D^2 is better approximated by a chi-square pdf with $K-1-r$ degrees of freedom, where r denotes the number of independent parameters estimated from x to fully specify the postulated pdf; thus, in effect, each estimated parameter decreases the degrees of freedom by one [13].

2.2. Classification based on Likelihood of Observations

In some cases, particularly in the analysis of DFT coefficient distributions to follow, we face the situation where a very limited number of observed data samples are available. In this case, we may expect rather inaccurate results if the procedure reviewed above is applied and alternatives are therefore of interest.

The procedure presented here assumes that any observed sample sequence x consists of iid samples drawn from one of the four candidate pdf types c_1 through c_4. Thus, we interpret our problem as a classification problem; the input space has been divided into four disjoint regions containing realizations from each pdf type, and given an observed sequence x we must decide to which region it belongs.

Let $P(c_i|x)$ denote the probability of pdf type i given the observed sequence x. In order to compute $P(c_i|x)$ we apply Bayes formula, and express $P(c_i|x)$ as

$$P(c_i|x) = \frac{f(x|c_i)P(c_i)}{f(x)}, \qquad i = 1, \dots, 4, \qquad (6)$$

where $f(x|c_i)$ is the class-conditional probability density, $P(c_i)$ denotes the a priori probabilities of the different pdf types, and $f(x)$ denotes the pdf of the observed data sequence. For a given c_i, the pdf type of $f(x|c_i)$ is given, and its parameters are estimated from x using maximum likelihood techniques. Subsequently, we compute the likelihood $f(x|c_i)$. As is often done when no specific a priori information is available [14], we assume uniform prior probabilities, $P(c_i) = 1/4, i = 1, \dots, 4$.

Having computed the values $P(c_i|x), i = 1, \dots, 4$, we assign the observation x to the pdf type c_i for which $P(c_i|x)$ is maximum. We note that this classification strategy is identical to *Bayes decision rule* and leads to a minimum error-rate classification (assuming that all decision errors are equally costly) [14].

2.3. Classification based on Moment Values

As an alternative to the previous method, we present here a classification method based on an estimate of the following quantity

$$M = \frac{E\{|x_n|\}}{\sqrt{E\{x_n^2\}}}, \qquad (7)$$

where $E\{\cdot\}$ denotes the statistical expectation operator, and x_n is a (scalar) data sample. The quantity M takes on different values for different underlying pdf types, but is independent of the actual pdf parameters. Thus, M provides a 'finger print' of

	c_1	c_2	c_3	c_4 ($\nu = 0.44$)
M	$\sqrt{2/\pi}$	$1/\sqrt{2}$	$1/\sqrt{3}$	0.5104

Table 1: Nominal values of M for different pdfs

the underlying pdf type. Indeed, it can be shown that M has the values given in Table 1 for different underlying pdf types.
The basic idea of our approach is to estimate M from the available data, and compare the estimate with the values of Table 1 to determine which of the candidate pdfs provides the best fit to the data. A similar approach was taken in [7], where visual comparison of estimated and expected M values was used to determine the underlying pdf type. Although visual inspection is certainly useful to get a first indication as to which pdf produced the observed data, we apply here a Bayesian classification method.
Let $P(c_i|\hat{M})$ denote the probability that the observed data sequence was produced according to the pdf c_i, for a given estimate $\hat{M}$ of M. Using Bayes formula, we express $P(c_i|\hat{M})$ as

$$P(c_i|\hat{M}) = \frac{f(\hat{M}|c_i)P(c_i)}{f(\hat{M})}, \qquad i = 1, \ldots, 4, \qquad (8)$$

where $f(\hat{M}|c_i)$ is the class-conditional probability density, $P(c_i)$ denotes the a priori probabilities of the different classes, and $f(\hat{M})$ denotes the pdf of the quantity $\hat{M}$.
We estimate the densities $f(\hat{M}|c_i)$ off-line from histograms of $\hat{M}$ values, computed from a large number of iid data sequences drawn according to c_i. It turns out that these estimated conditional densities are dependent on the number of samples N in the data sequence used for estimating $\hat{M}$, so we repeat the procedure for different values of N. As in the previous section we assume uniform prior probabilities $P(c_i) = 1/4$, and as before the optimal decision rule is to assign the observed sequence x to the class c_i for which $P(c_i|\hat{M})$ is largest [14].

3. DISTRIBUTION OF TIME-DOMAIN SAMPLES

We study the distribution of time domain speech samples based on speech material from the Timit database [15]. More specifically, the results reported here are based on a subset of the Timit data base containing approximately 20 minutes of speech, spoken by 12 female and 25 male speakers, downsampled to f_s =8 kHz. First, the speech sentences are segmented into consecutive segments of N samples[4], for varying values of N, taken with an overlap of 50% . For each segment and candidate pdf type, the goodness-of-fit value D^2 is computed. Fig. 1a shows the average D^2 value, for different segment lengths N. We used a value of $K = 20$ in Eq. (5) resulting in a threshold $t_{\alpha,\chi^2} = 34.81$ for $\alpha = 1$ %. With this threshold, an analysis of Fig. 1a suggests that the null hypothesis is accepted for all segment lengths shorther than 200 ms, independent of the hypothesized distribution type. However, although the chi-square test does not allow a clear identification of 'the best model', the results do suggest that the Laplacian model is advantageous for segment lengths of 20–50 ms. In contrast to the results in Fig. 1a, we now restrict the

[4]For all tests, we discard segments containing non-speech related sounds.These sounds are typically located before or after the actual speech excerpt and are annotated with the symbol 'h#' in the Timit data base.

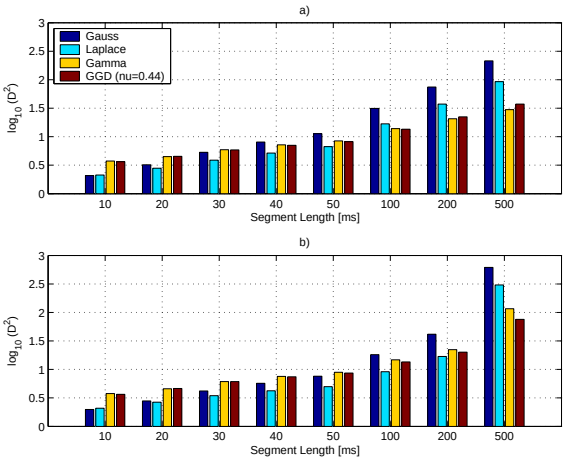

Figure 1: *Average D^2 values for different segment lengths when segments may fall across speech sound boundaries (top) and for segments within speech sound boundaries (bottom).*

segments to occur within the hand-labeled speech sound boundaries provided with the Timit data base. We do this to ensure that the signal segment under analysis can be assumed roughly stationary. Fig. 1b shows resulting average D^2 values. We see that the D^2 values, especially for the Gaussian and Laplacian distribution have decreased. In particular, the Laplacian model now provides the best fit for segment lengths 20–200 ms.
We then analyze in further detail the results shown in Fig. 1b for the segment length 20 ms. We divide the analyzed segments into broad speech classes (stops, glides, fricatives, nasals, and two groups of vowels) using the hand-labeled classifications provided with the Timit database, and then compute the average D^2 values of the pdf models within these classes. The result of this procedure is shown in Figs. 2a and 2b, for male and female speakers, respectively. We see that, generally, the speech classes containing fricatives, nasals, and glides appear to be better represented with the Gaussian model, while stop consonants are much better modeled with the Laplacian model, making the Laplacian model slightly better than the Gaussian on average.

4. DISTRIBUTION OF DFT COEFFICIENTS

This section aims at analyzing the distribution of Discrete Fourier Transform (DFT) coefficients. The study is of interest in for example speech enhancement contexts, where DFT coefficients have been modeled as iid random variables drawn from a Gaussian, e.g. [11], a Laplacian [9], and a Gamma distribution [8]. In this study, we fix the analysis segment length to 32 ms ($N = 512$ at a sampling rate of f_s = 16 kHz), taken with an overlap of 50%. We restrict the analysis segments to be taken within speech sound boundaries. The segments are extracted using a Hamming window and transformed to the Fourier domain using an N'th order FFT. We assume that the real and imaginary parts of the complex DFT coefficent are iid random variables, a standard assumption in most DFT based speech enhancement systems [11, 8, 9]. In this way we get within a given speech sound $2N_s$ real-valued observations per DFT channel, where N_s denotes the number of segments within the speech sound in question.
For each DFT channel we compute the probabilities $P(c_i|x), i =$

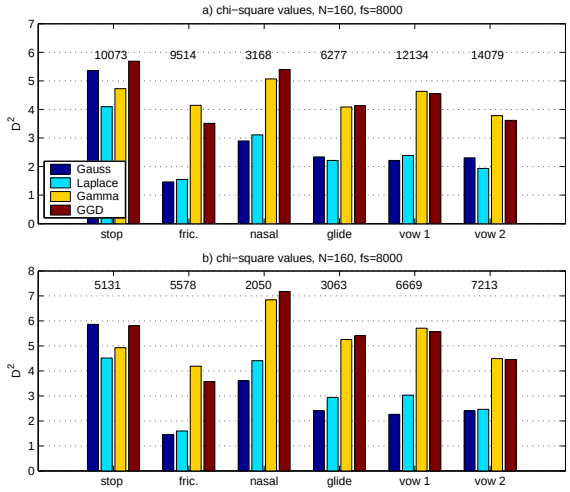

Figure 2: *Average D^2 values across broad speech classes for $N = 160$, f_s=8 kHz, for male speakers (top) and female speakers (bottom). Number of segments within each speech class is shown above bars.*

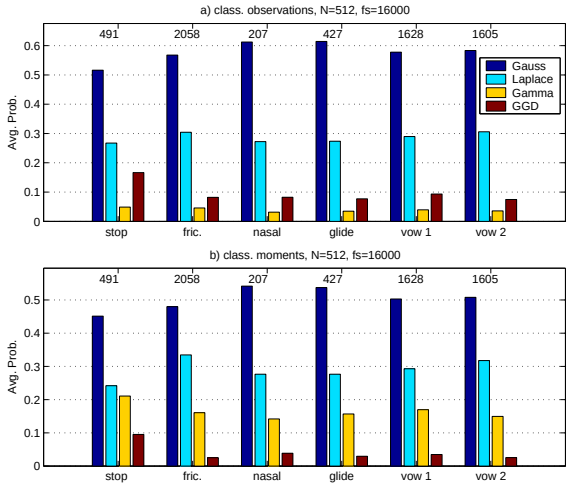

Figure 3: *Percentages of DFT bins classified as Gaussian, Laplacian, Gamma, and GGD using the log-likelihood of the observations (top) and the moment values, Eq. (7) (bottom). Number of occurrences of speech sounds shown above bars.*

$1, \ldots, 4$, see Eq. (6) and assign each channel to the pdf type with the largest probability[5]. To present these results in a managable way, we count the percentages of bins assigned to each pdf type across DFT channels and across all occurences of the speech sound in question. In this way we get an estimate of the probability for each pdf type c_i for each speech sound. The result of this procedure is shown in Fig. 3a. A similar procedure for the probabilities $P(c_i|\hat{M}), i = 1, \ldots, 4$, in Eq. (8), leads to the results in Fig. 3b.

We see that DFT coefficients are generally classified as Gaussian, for all speech classes. This result appear to be in contradiction with the work in [8, 9], which suggests that a Laplacian or Gamma distribution offers a better fit. We believe the difference may be explained by the fact that [8, 9] base their results on DFT coefficents collected across longer time spans and several speakers, while the results presented here are based on DFT coefficients collected within a single speech sound, and thus across a much shorter time interval.

5. ACKNOWLEDGEMENT

This research was partly supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

6. REFERENCES

[1] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*, Elsevier Science B. V., 1995.

[2] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc., 1993.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, October 1993.

[5] In this study we omit speech sound realizations for which $N_s < 5$

[4] W. B. Davenport, "An experimental study of speech-wave probability distributions," *J. Acoust. Soc. Am.*, vol. 24, pp. 390–399, July 1952.

[5] M. D. Paez and T. H. Glisson, "Minimum mean-square error quantization in speech," *IEEE Trans. Comm.*, vol. com-20, pp. 225–230, April 1972.

[6] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.

[7] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Lett.*, vol. 10, no. 7, pp. 204–207, July 2003.

[8] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Orlando, Florida, 2002, vol. I, pp. 253–256.

[9] R. Martin and C. Breithaupt, "Speech enhancement in the dft domain using laplacian speech priors," in *Int. Workshop on Acoustics, Echo and Noise Control (IWAENC)*, Kyoto, Japan, September 2003, pp. 87–90.

[10] D. Brillinger, *Time Series: Data Analysis and Theory*, Holden-Day, 1981.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[12] J.-H. Chang, J.-W. Shin, and N. S. Kim, "Likelihood ratio test with complex laplacian model for voice activity detection," in *Eurospeech*, Geneva, 2003, pp. 1065–1068.

[13] A. O. Allen, *Probability, Statistics, and Queueing Theory*, Academic Press, Inc., New York, 1978.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, Inc., 2 edition, 2001.

[15] DARPA, "Timit, acoustic-phonetic continuous speech corpus," NIST Speech Disc 1-1.1, October 1990.