Statistical Modeling of the Speech Signal

Ivan Tashev, Alex Acero
Speech Technology Group
Microsoft Research
Redmond, WA 98052, USA
{ivantash, alexac}@microsoft.com

***Abstract*—The Gaussian distribution is the most commonly used statistical model of the speech signal. In this paper we propose more general statistical model for the distributions of the real and imaginary parts of the speech signal DFT coefficients and their magnitudes. Based on experimental measurements with the TIMIT database we have shown that the Generalized Gaussian Distribution holds well across frequency and audio frame size. A Weibull distribution is proposed to model the statistical behavior of the speech signal amplitude in the frequency domain. Estimation of the distribution parameters from experimental measurements corresponds well to the distribution of the real and imaginary parts. We propose and evaluate several statistical models of various complexities. Overall these statistical models fit the actual measurements with a Jensen-Shannon divergence below 0.0012 for real and imaginary parts and below 0.003 for magnitudes. The results presented in this paper are applicable for improving speech processing algorithms based on statistical signal processing.**

Keywords-speech statistical model, generalized Gaussian distribution, Weibull distribution.

I. INTRODUCTION

Statistical models of the speech and noise signals play an important role in single channel speech enhancement, multichannel speech processing for microphone arrays, voice activity detectors, speech compression, and in many other statistical signal processing algorithms. The real and imaginary parts of the speech signal spectrum coefficients are very often modeled as independent and identically distributed zero mean Gaussian variables, which is motivated by the central limit theorem. Modeling the speech signal DFT coefficients as zero mean Gaussian processes is utilized in the derivation of most noise suppression algorithms: Weiner [1], short term spectral minimum mean square estimators [2], and short term spectral log-minimum mean square estimators [3]. As most of the suppression based algorithms actually estimate only the magnitude and take the phase from the noise corrupted signal, it is important to observe that under this assumption the speech signal magnitudes have Rayleigh distribution. In [4] the derived set of suppression rules assumes Laplace and Gamma distributions of the speech signal spectrum. Later measurements [5] concluded that the speech signal distribution in time domain is frame size dependent and has a super-Gaussian distribution, i.e. its PDF is more "peaky" than the bell-shaped Gaussian PDF. A step further into using the super-Gaussian distribution of the speech signal is presented in [6], where the author fits several potential distributions to measured histograms of speech signal DFT coefficients and magnitudes and derives suppression rules for the generic distribution of the speech and noise signals. The Gaussian PDF carries the least information as it has the highest entropy. Thus the use of any other PDF is attractive as it carries more information. In [7] minimizing the entropy is used as an adaptation criterion for an adaptive beamformer.

The goal of this paper is to provide more precise measurements of the speech signal PDF as a function of the frequency and the frame size. We use Generalized Gaussian Distribution to model the distribution of the real and imaginary parts of the speech signal DFT coefficients, and propose using the Weibull distribution to model the distribution of the speech signal magnitudes in frequency domain. Both distributions are dual parameters and in addition to variance have shape parameters which determine the "peakiness" of the distribution. We measure the shape parameters using a clean speech corpus and propose four models with various complexities.

II. MODELING

Let us assume that we have speech signal with an average power spectrum $\lambda_s(k)=E\left[|X_k|^2\right]$. Further in this paper we will omit the frequency bin index (k) wherever it is possible.

A. Real and imaginary parts of the DFT coefficients

Since the DFT coefficients are complex numbers we can assume that the real and imaginary parts are zero mean, independent, and identically distributed, i.e. they have the same variance $\lambda_s/2$. A flexible model for the distribution is Generalized Gaussian Distribution (GGD), defined in [8] and [9] with the PDF:

$$p(x\mid\mu,\alpha,\beta)=\frac{\beta}{2\alpha\Gamma(1/\beta)}\exp\left(-\left(\frac{|x-\mu|}{\alpha}\right)^{\beta}\right). \qquad (1)$$

Here α is the scale parameter and β is the shape parameter. The mean is μ, the variance is $\alpha^2\Gamma(3/\beta)/\Gamma(1/\beta)$, and $\Gamma(\cdot)$ denotes the Gamma function. When $\beta=2$ GGD con-verges to a Gaussian, and at $\beta=1$ the PDF converges to a Laplace distribution. For a zero mean distribution the scale parameter can be estimated from the average power spectrum:

$$\alpha=\sqrt{\frac{\lambda_s}{2}\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}\,. \qquad (2)$$

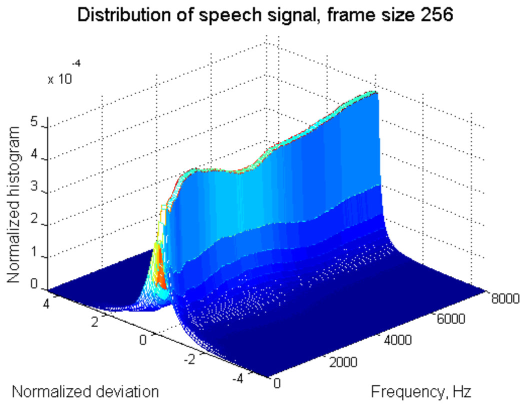

Figure 1. Distribution of the real and imaginary parts of the analyzed speech signal from TIMIT.

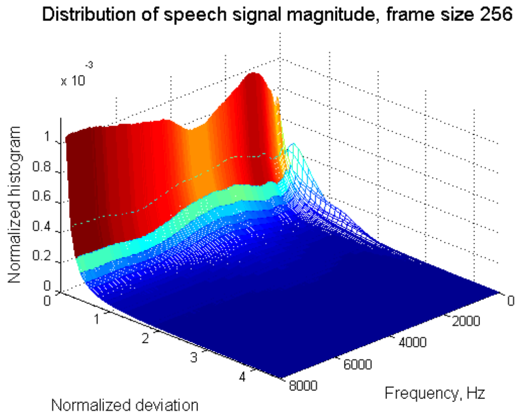

Figure 2. Speech signal magnitudes from analyzed speech signal in TIMIT.

B. *Magnitudes of the DFT coefficients*

The distribution of the signal magnitudes is modeled with the powerful and flexible Weibull distribution [10], given by:

$$p(x \mid \chi, \delta) = \frac{\delta}{\chi}\left(\frac{x}{\chi}\right)^{\delta-1} \exp\left(-\left(\frac{x}{\chi}\right)^{\delta}\right), \quad x \geq 0 \qquad (3)$$

where χ is the scale parameter and δ is the shape parameter. The mean is $\chi\Gamma(1+1/\delta)$ and $\chi^2\left[\Gamma(1+2/\delta)-\Gamma^2(1+1/\delta)\right]$ is the variance. When $\delta = 2$ the Weibull distribution converges to a Rayleigh, and at $\delta = 1$ it converges to an exponential distribution. The scale parameter can be estimated from the average power spectrum:

$$\chi = \sqrt{\frac{\lambda_s}{\Gamma(1+2/\delta)}}. \qquad (4)$$

C. *Fitting criterion*

The actual distributions of the speech signal can be measured as a histogram of the real and imaginary parts, or as a histogram of the magnitudes. The scale parameters α or χ can be estimated from the signal power spectrum and used for normalization of the histograms. The distance between the two probability distributions is given by the Jensen-Shannon divergence [11], which is a symmetrized and smoothed version of Kullback-Leibler divergence [12] $D_{KL}(p\|q)$:

$$D_{JS}(p\|q) = \frac{1}{2}\left(D_{KL}(p\|m) + D_{KL}(q\|m)\right), \; m = \frac{1}{2}(p+q). \qquad (5)$$

The Kullback-Leibler divergence is defined as:

$$D_{KL}(p \,\|\, q) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} \qquad (6)$$

and measures the expected number of extra bits, required to code samples from q when using a code based on p, rather than using a code based on q. Lower Jensen-Shannon divergence (D_{JS}) indicates a better fit of the model to the measured histogram. We are going to estimate the shape parameter β as:

$$\beta_{opt} = \arg\min_{\beta}\left(D_{JS}(p \,\|\, q)\right). \qquad (7)$$

Here p is the measured distribution (the histogram) and q is the distribution model, which depends on β. The same approach can be used to estimate the shape parameter δ for fitting the magnitudes distribution.

III. Experimental Results

A. *Speech corpus*

The clean speech corpus known as TIMIT [13] was selected for measuring the distribution of the speech signals. TIMIT contains 6300 utterances of clean speech sampled at 16 kHz with 16 bits precision. The corpus is conveniently split on training (~4 hours) and test (~1.5 hours) sets.

B. *Processing of the training data*

The set of frame sizes is selected to cover the most frequently used frame sizes of 32, 40, 64, 80, 128, 160, 256, 320, 640, and 1024 samples. These numbers are the length of a half size complex vector of the DFT coefficients, i.e. the frame durations are from 4 to 128 milliseconds. Audio frames are with 50% overlapping and weighted with Hann window – typical settings for most of the audio processing pipelines.

Each file is converted to frequency domain with a given frame size and an energy based voice activity detector is used to select the frames with a speech signal. For each frequency bin the speech frames are processed as follows:

- extract the real and imaginary parts and combine them;
- compute the variance for each frequency bin, normalize;
- build the histogram in the range from -4.6 to +4.6 times the deviation with a step of 0.1, leading to 93 bins.

The same normalization and histograming process is repeated with the magnitudes of the DFT coefficients, except that the scale is from 0 to 4.6 times the deviation, leading to 47 magnitude bins. Overall for each frame size we have 400 million data points for the training set and 137 million data points for the test set.

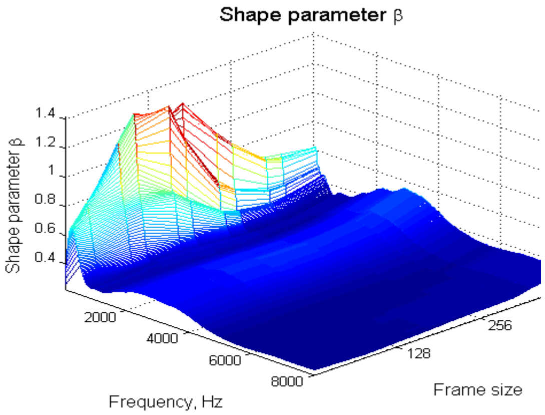

Figure 3. The shape parameter β as a function of the frequency and frame size.

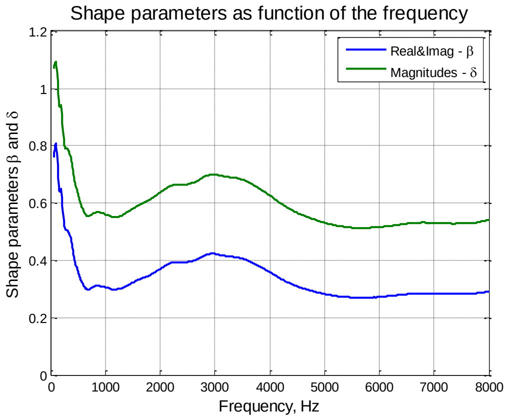

Figure 5. Shape parameters β and δ as a function of the frequency.

The histograms from all files in the training set are combined together and as a result for each frame size K we have a matrix with dimensions Kx93 for the real and imaginary parts and Kx47 for the magnitudes. The measured distribution of the speech signal for a frame size of 256 samples is shown in Figure 1 and for magnitudes in Figure 2. It is well visible that the distribution of the speech signal is super-Gaussian, which leads to a magnitude distribution closer to an exponential rather than to Rayleigh. In addition the “peakiness” of the distribution is quite uniform in the middle and upper frequency range, but opens slightly towards the lowest part of the frequency band.

For each frequency bin and frame size we computed the shape parameters β and δ using (7), which results in the best fit of GGD or Weibull distributions respectively. We used a one dimensional version of the steepest gradient descent algorithm, but practically any one dimensional optimization algorithm can find the best solution. The histogram bins with zero values are excluded from the fitting process. The reason for this is that even in the speech frames most of the frequency bins do not contain a speech signal. As TIMIT is a very clean speech corpus these frequency bins contain close to zero values and fall into the zero histogram bins. The measured function $\beta(K,f)$ for the speech signal is shown in Figure 3. It confirms the observation from the previous paragraph. The shape of the measured function $\delta(K,f)$ is quite similar, as these two functions describe the same process.

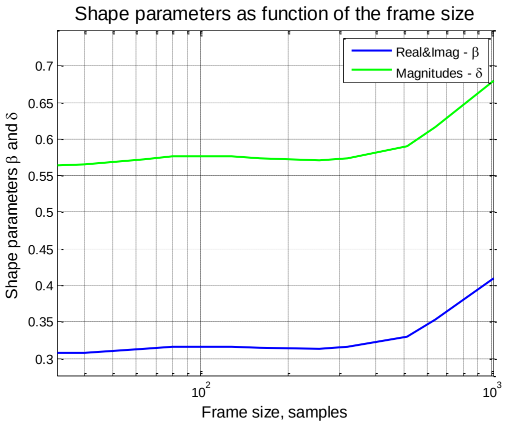

Figure 4. Shape parameters β and δ as a function of the frame size.

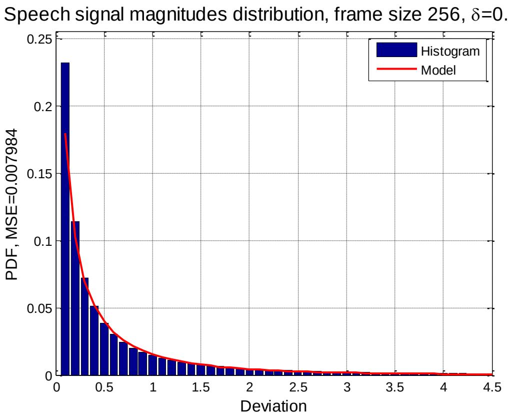

Figure 6. Speech signal magnitude and fitted distribution.

By combining all frequency bins we can compute the shape parameters as a function only of the frame size $\beta(K)$ and $\delta(K)$, shown in Figure 4. By averaging across frequencies we derive the frame size independent measurements of the shape parameter as a function of the frequency $\beta(f)$ and $\delta(f)$, shown in Figure 5. Finally, by combining all points we can measure single shape parameter, independent of frame size and frequency, which are $\beta = 0.327$ and $\delta = 0.588$.

C. *Evaluation with the test data*

We have measured four models of the speech signal distributions with various complexity, starting with single numbers, and ending with a two dimensional function of the shape parameters. For evaluation of these four models we used the test set in the TIMIT database. It was processed in the same way as the training set, and the resulting histograms are evaluated with distribution models, derived in the previous paragraph using Jensen-Shannon divergence as evaluation criterion. As a base line we use Gaussian and Laplace distributions for real and imaginary parts. The summary of the results is shown in Table 1. Table 2 presents the results for frame size dependent estimations. Figure 6 shows the histogram and the modeling PDF for magnitudes and frame size 256 samples.

TABLE I. RESULTS SUMMARY

PDF Fitting	Real& Imag		Magnitudes	
	β	D_{JS}	δ	D_{JS}
Gaussian	2.000	0.12242	2.000	0.23244
Laplace	1.000	0.06329	1.460	0.13430
One shape parameter	0.327	0.00164	0.588	0.00346
Frame size dependent		0.00115		0.00297
Frequency dependent		0.00271		0.00604
Both freq. and frame size		0.00185		0.00505

IV. DISCUSSION

The results in Table 1 indicate that the more precise models presented in this paper outperform modeling of the speech signal with Gaussian and Laplace distributions. Using the same shape parameter for all frame sizes and frequencies still reduces the Jensen-Shannon divergence by more than two magnitudes. Increasing the complexity of the model reduces the divergence further, but using shape parameter as a function of two parameters is marginally better than the shape parameter as a function of only the frequency. The lowest average divergence we achieved with the frame size dependent model. We conclude that the dependency of the shape parameters on the frame size is stronger than the dependency on the frequency. By increasing the model complexity to depend on both frame size and frequency the precision decreases, which is an indication of overtraining to the data in the training corpus. Using the frame size dependent models, shown in Table 2, is straightforward and trivial. The results in Table 1 are averaged using the same weight for all frequency bins.

Similar measurements with various noise signals, not presented in this paper, confirmed that the best fit for the real and imaginary parts distribution of the noise spectra is a Gaussian distribution, which corresponds to a Rayleigh distribution of the magnitudes.

V. CONCLUSIONS AND FUTURE WORK

In this paper we presented four models of the distribution of the real and imaginary parts of the speech signal DFT coefficients and four models of the distribution of the magnitudes. The proposed models use the Generalized Gaussian and the Weibull distributions to describe the statistical parameters of the speech signal. They are more precise, compared to the Gaussian and Laplace distributions widely used today. Potential future steps are in two directions. The first is further improvement of the models. This includes more precise measurements and using languages other than English. While the speech production apparatus is the same for all humans, it still has to be proven how well these models hold for other languages. The second direction is using these more precise models in speech enhancement, microphone array processing, speech compression, etc. While in statistical signal processing better distribution models lead to better results in general, the improvement in sound quality due to these more precise models still has to be verified in real life applications.

TABLE II. FRAME SIZE RESULTS

Frame size	Real& Imag		Magnitudes	
	β	D_{JS}	δ	D_{JS}
32	0.307	0.00136	0.564	0.00348
40	0.308	0.00147	0.565	0.00311
64	0.313	0.00157	0.573	0.00255
80	0.316	0.00156	0.577	0.00241
128	0.316	0.00129	0.576	0.00242
160	0.314	0.00110	0.573	0.00255
256	0.313	0.00089	0.571	0.00312
320	0.316	0.00086	0.574	0.00338
512	0.330	0.00085	0.591	0.00365
640	0.353	0.00083	0.617	0.00338
1024	0.410	0.00093	0.680	0.00268

REFERENCES

[1] Wiener, N. Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications, MIT Press, Cambridge, MA, USA, 1949.

[2] Ephraim, Y.; Malah, D. “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator”. IEEE Trans. on Acoust., Speech, and Signal Processing, vol. ASSP-32, No. 6, December 1984, pp. 1109-1121.

[3] Ephraim Y.; Malah, D. “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator”. IEEE Trans. on Acoust., Speech, and Signal Processing, vol. ASSP-33, No. 2, pp. 443–445, Apr. 1985.

[4] Martin, R. “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors”. Proc. of IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP), pp. 253-256, Orlando, Florida, USA, 2002.

[5] Gazor, S.; Zhang, W. “Speech Probability Distribution”. IEEE Signal Processing Letters, vol. 10, No. 7, pp204-207, July 2003.

[6] Lotter, T. “Single- and Multi-Microphone Spectral Amplitude Estimation Using a Super-Gaussian Speech Model”, in J. Benesty, S. Makino, and J. Chen [Ed], Speech Enhancement, pp. 67-95, Springer-Verlag Berlin Heidelberg, 2005.

[7] Kumatani, K.; McDonough, J.; Klakow, D.; Garner, P.N.; Weifeng Li. "Adaptive Beamforming with a Maximum Negentropy Criterion". Proc. of Hands-Free Speech Communication and Microphone Arrays, HSCMA 2008. Trento, Italy, May 2008.

[8] Varanasi, M.K.; Aazhang, B. "Parametric generalized Gaussian density estimation". Journal of the Acoustical Society of America 86 (4): 1404–1415, October 1989.

[9] Nadarajah, S. "A generalized normal distribution". Journal of Applied Statistics 32 (7): 685–694, September 2005.

[10] Weibull, W. "A statistical distribution function of wide applicability" J. Appl. Mech.-Trans. ASME 18(3), 293-297, 1951.

[11] T. Cover, J. Thomas. Elements of Information Theory. Wiley, New York, NY, 1991.

[12] Kullback, S.; Leibler, R. “On information and sufficiency”, Annals of Mathematical Statistics 22:79-86, 1951.

[13] John S. Garofolo, et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus" Linguistic Data Consortium, Philadelphia, 1993.