INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Estimation of the Probability Distribution of Spectral Fine Structure in the Speech Source

Tom Bäckström

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
`first.lastname@aalto.fi`

Abstract

The efficiency of many speech processing methods rely on accurate modeling of the distribution of the signal spectrum and a majority of prior works suggest that the spectral components follow the Laplace distribution. To improve the probability distribution models based on our knowledge of speech source modeling, we argue that the model should in fact be a multiplicative mixture model, including terms for voiced and unvoiced utterances. While prior works have applied Gaussian mixture models, we demonstrate that a mixture of generalized Gaussian models more accurately follows the observations. The proposed estimation method is based on measuring the ratio of L_p-norms between spectral bands. Such ratios follow the Beta-distribution when the input signal is generalized Gaussian, whereby the estimated parameters can be used to determine the underlying parameters of the mixture of generalized Gaussian distributions.

Index Terms: probability distribution mixture models, speech production modeling

1. Introduction

Modeling the probability distribution is central in many approaches to speech coding, enhancement, synthesis and recognition [1, 2, 3, 4]. For example, in coding of the short-time spectra of speech signals, the Laplacian has been shown to be both accurate, but also computationally simple to implement [5, 6]. Similarly, compensating for the mismatch between Gaussian and Laplacian distributions can be used to improve performance in speaker recognition [7].

Selection of the best probability distribution model is therefore important for the efficiency of the algorithms, whereby many have investigated the issue [8, 9, 10]. In general, Gaussian models and mixture models thereof are appealing due to their analytic and computational properties, but practice have shown that in many cases the Laplacian distribution is a more accurate model of many representations of speech signals.

In this work, we wish to improve the above models of the probability distributions by including aspects of speech production models. Speech signals can be efficiently modeled with the classic source-filter model, which consists of a sum of periodic and noisy excitations modulated by an IIR-filter [1, 2]. Moreover, the overall signal can be modulated by a global gain factor. The distribution of each component of the speech production model should then be separately modeled.

Note that the signal intensity is rather straightforward to model, as it is a relatively slowly changing signal. Moreover, the signal intensity depends on both the speech source, but also the physical properties of the recording setup. Modeling the distribution of signal intensity is therefore not particularly interesting in this context.

The spectral envelope of the signal is an important feature of the speech signal, since it carries the identity of many phoneme classes, such as vowels. The distribution of the envelope parameters depends highly on the representation used, such as Mel-frequency cepstral coefficients, linear predictive models or distribution coding [2, 1, 11]. Modeling of the spectral envelope is however not our focus here, whereby we will leave it for further study.

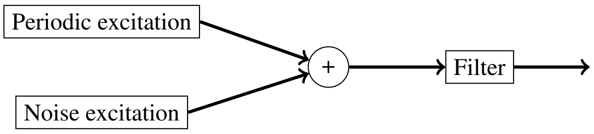

Figure 1: *Source-filter modeling of speech.*

The remaining part of the speech production model is the speech excitation, which corresponds to the spectral fine-structure. That is, since the gross-shape of the spectrum is modeled by the spectral envelope, the remaining parts are the micro-structures of the spectrum. It consists of two parts, the voiced and unvoiced excitations. Our hypothesis is therefore that we will see a mixture of at least two distributions, corresponding to each of the phonation types. Note however that all speech sounds have a noisy component, whereby we do not expect to see a clear division into two categories, but an extensive overlap between classes of phonemes.

To analyze spectral fine-structure in isolation of the effects from overall intensity as well as the spectral envelope, we need to normalize the signal carefully. It is obvious that, for example, variations in signal intensity will have a big influence on the probability distribution of spectral components across time. Similarly, changes in the spectral envelope will affect the components across the frequency range. We can remove the effect of both the signal gain and the spectral envelope by normalizing the spectrum of a frame bandwise.

Normalization of signal has, however, a significant impact on the distribution of the spectral coefficients. Namely, normalization limits the range of the coefficients to a finite field. If we would assume that the actual distribution is Gaussian or Laplacian, and apply normalization, then the signal would no longer be Gaussian or Laplacian. We shall see, in the following sections, that the normalized amplitude of general Gaussian signals follow the Beta distribution.

The main contributions of this paper are i) development of methods for estimating the distribution of normalized signals, when the signal is assumed to follow the generalized Gaussian signal and ii) application of the methods in analysis of the mixture of distributions in spectral fine structure of speech signals.

2. Speech Production Modeling

Speech signals can be efficiently and simply modeled with the source-filter model [2], where the speech signal is modeled in

http://dx.doi.org/10.21437/Interspeech.2017-389

three parts; periodic and noisy excitations as well as a filter which modulates the sum of the two excitations (see Fig. 1).

If x_n and v_n are the periodic and noisy excitations, and h_n is the impulse response of the filter, then the output time signal is defined as

$$s_n = h_n * (x_n + v_n). \tag{1}$$

By defining the Z-transforms $X(z)$, $V(z)$, $H(z)$, and $S(z)$, of x_n, v_n, h_n and s_n, respectively, we can equivalently write in the frequency domain

$$S(z) = H(z)\left[X(z) + V(z)\right]. \tag{2}$$

We can thus see that, when analyzing the distribution of the output signal $S(z)$, it will be a product between $H(z)$ and the sum $X(z)$ and $V(z)$.

Direct analysis of multiplicative models is difficult, but since the spectral envelope, corresponding to the filter $H(z)$ is a smooth shape, we can use that to our advantage. Namely, if we analyze a sufficiently small neighborhood around z, then the filter magnitude will be roughly constant, $|H(z+\epsilon)| \approx |H(z)|$. By normalizing $S(z)$ with an approximation $|\hat{H}(z)|$ of $H(z)$, we thus gain access to the excitations

$$\frac{|S(z)|^2}{|H(z)|^2} \approx |X(z) + V(z)|^2 \approx |X(z)|^2 + |V(z)|^2, \tag{3}$$

where we have assumed that $X(z)$ and $V(z)$ are independent. By local normalization of the spectrum, we thus obtain access to the excitation, whereby we can analyze the distribution of the mixture components.

3. Probability distributions of normalized signals

Prior works have found that speech signals follow a distribution similar to the Laplacian and Gaussian distributions. Our purpose is to generalize these distributions using the generalized Gaussian distribution. Moreover, we need to analyze the distributions for a locally normalized signal, corresponding to the excitation.

The generalized Gaussian distribution is defined for a sample ξ as [12]

$$f(\xi) = \frac{p}{2\rho\Gamma(1/p)} \exp\left(-\frac{|\xi|^p}{\rho^p}\right), \tag{4}$$

where ρ is a scaling factor and $\sigma^2 = \rho^2\frac{\Gamma(1/p)}{\Gamma(3/p)}$ is the variance. Notably, with $p = 2$ and $p = 1$ we obtain the normal and Laplacian distributions, respectively.

Let us form the distribution of the pth power of ξ, readily obtained by a change of variable $\tau = 2|\xi|^p\rho^{-p}$, whereby

$$f(\tau) = f(\xi)\frac{d\tau}{d\xi} \propto \tau^{\frac{1}{p}-1}e^{-\frac{\tau}{2}} \sim \chi^2\left(\frac{2}{p}\right). \tag{5}$$

This means that τ follows the Chi-squared distribution with $k = \frac{2}{p}$ degrees of freedom.

Secondly, we can define a vector of $x = [\xi_1, \dots, \xi_N]^T$ of uncorrelated generalized Gaussian samples, and define its scaled L_p-norm as

$$\phi_x = \frac{2\|x\|_p^p}{\rho^p} = \frac{2\|x\|_p^p}{\sigma^p}\left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{p/2}. \tag{6}$$

Since ϕ is then a sum of N Chi-squared distributed variables with $k = \frac{2}{p}$ degrees of freedom each, then ϕ will be Chi-squared distributed with $k = \frac{2N}{p}$ degrees of freedom.

The final step is to derive the distribution of normalized vectors. For that purpose we first define the norm-ratio of two vectors $x \in \mathbb{R}^{N\times 1}$ and $y \in \mathbb{R}^{M\times 1}$ as

$$\lambda = \frac{\|x\|_p^p}{\|x\|_p^p + \|y\|_p^p} = \frac{\phi_x}{\phi_x + \phi_y}. \tag{7}$$

If x and y follow the generalized Gaussian distribution with the same parameters p and ρ, then ϕ_x and ϕ_y will follow the Chi-squared distribution with $\frac{2N}{p}$ and $\frac{2M}{p}$ degrees of freedom, respectively. The ratio above therefore (for details, see e.g. [13]) follows the Beta-distribution, $\lambda \sim \text{Beta}\left(\frac{N}{p},\, \frac{M}{p}\right)$, that is,

$$f(\lambda) \propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}, \tag{8}$$

where $\alpha = \frac{N}{p}$ and $\beta = \frac{M}{p}$.

The main result is thus that, if the norm-ratio of a signal (as specified in Eq. 7) follows the Beta-distribution with the parameters above, then the underlying signal *is congruent* with the generalized Gaussian distribution. However, since we have here only observed the absolute value of the signal, we must additionally check that the distribution of the signal is symmetric around zero, to be able to claim that the signal *follows* the generalized Gaussian distribution.

Additionally, we can derive the distribution of the normalized, generalized Gaussian signal. We thus want to determine the distribution of a sample ξ_1 within a vector x which was normalized by $\|x\|_p$. By setting $N = 1$ in Eq. 7, we have $\lambda = |\xi_1|^p/\|x\|_p^p$ and we can define

$$\eta = \frac{\xi_1}{\|x\|_p}. \tag{9}$$

Since $\lambda = |\eta|^p$, we obtain the distribution of η as

$$\begin{aligned} f(\eta) = f(\lambda)\frac{d\eta}{d\lambda} &\propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}|\eta|^{p-1} \\ &\propto (1-|\eta|^p)^{(N-p-1)/p}. \end{aligned} \tag{10}$$

We can readily see that η thus follows the Beta-distribution, with $\alpha = \beta = \frac{N-1}{p}$, which is scaled and translated to the range $\eta \in [-1, +1]$.

4. Experiments

To evaluate the performance of the proposed method for estimating parameters of probability distributions, as well as the fitness of said probability models on speech data, we studied the TIMIT corpus [14]. Using the labeling, we extracted all speech phonemes by discarding silence and pauses. The signal was then windowed into 20 ms segments, using half-sine windows. This windowing scheme was adopted to match modern speech codecs [15, 5], such that the results will be directly applicable in the design of future codecs [16]. Each window was further converted to the frequency domain with the discrete cosine transform, to obtain spectral representations of length 320 samples.

The spectrum of each frame was then split into bands of 8 samples, corresponding to a bandwidth of 200 Hz. We can assume that the spectral envelope is constant within such a band, whereby our assumption in Eq. 3 holds. The bands were further

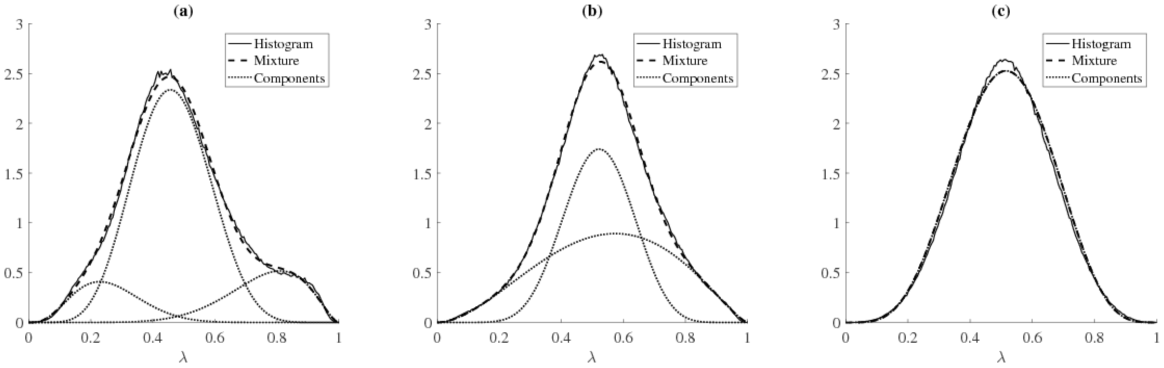

Figure 2: *Examples of the histograms of the norm-ratio λ_p from Eq. 11 for $p = 1$ in the bandwidths (a) 200 Hz to 400 Hz, (b) 1000 Hz to 1200 Hz and (c) 5000 Hz to 5200 Hz over all phonemes in the TIMIT corpus. The best fitting beta mixture model and its individual components are illustrated, such that we have (a) $K = 3$, (b) $K = 2$ and (c) $K = 1$ components.*

split in the middle, to a left x_L and a right x_R part, each of width 4 samples. The norm-ratios for each p, and over all bands and frames was then determined

$$\lambda_p = \frac{\|x_L\|_p^p}{\|x_L\|_p^p + \|x_R\|_p^p}. \tag{11}$$

According to our theory, this parameter should follow the Beta-distribution if the excitation follows the generalized Gaussian distribution.

Our hypothesis is that since the speech excitation contains two sources, we should be able to observe a mixture of at least two Beta-distributions. Moreover, our task is to determine the exponent p for each of the generalized Gaussians in the mixture.

To fit a mixture of Beta-distributions to the observations, we used an implementation of the standard expectation maximization (EM) algorithm. To improve convergence, we applied an exponentially decaying forgetting factor in the EM-iteration [17]. Parameters of each Beta-distribution was determined from a sample with the method of moments [18]. Examples of the performance of the fitting algorithm are presented in Fig. 2. Visual inspection verifies that the fit of the distributions is accurate.

To verify our hypothesis that the signal is a mixture of multiple excitations, we further estimate the parameters for different number K of Beta-distributions. We investigated informally the use of Akaike and Bayesian information criteria for selection of the model order, but in the end, the log-likelihood turned out to be a simple method with best match visual evaluation of model fitness. We evaluated the log-likelihood of from $K = 1$ to 3 and if the improvement in log-likelihood was less than 1 %, then the higher model order was discarded. The results are illustrated in Fig. 3.

We observe that the spectrum is, on the main range $p = 0.7$ to 2.0, split into three parts; for the high frequencies (above approx. 3.5 kHz), the speech signal has only one component, the mid-frequencies (0.6 kHz to 3.5 kHz) have mostly two components, whereas the low frequencies have a varying number of components. The mid-frequencies thus follow the original hypothesis of two components. At the high frequencies the harmonic structure is less pronounced, whereby it is natural that we see only one signal. Informal observations at the lower frequencies would suggest that three components would be the true number for all p, but that the expectation maximization algorithm did not converge properly.

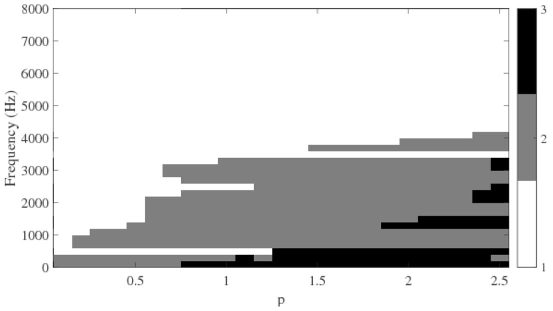

Figure 3: *Number of components identified in the mixture at different frequencies and for different values of p.*

The length of the vectors x_L and x_R are $N_L = 4$ and $N_R = 4$ respectively, whereby the parameters of the Beta-distributions would be, in theory, $\alpha = \frac{N_L}{p}$ and $\beta = \frac{N_R}{p}$. Conversely, given the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$, we can determine the corresponding vector lengths as $\hat{N}_L = \hat{\alpha} p$ and $\hat{N}_R = \hat{\beta} p$ and the total length is then $\hat{N} = \hat{N}_L + \hat{N}_R$. If the total length coincides with the true length $N = 8$, then that supports the hypothesis that the p value used in normalization in Eq. 11 matches the true distribution of the signal.

The main interest of these experiments was to determine the power p, for the different excitations, such that we can determine whether they are more like Laplacian ($p = 1$), Gaussian ($p = 2$) or something else. For that purpose, we evaluated the models with p in the range 0.1 to 2.5 in steps of 0.1 and estimated the corresponding values $\hat{N}$. Fig. 4 illustrates the estimated values $\hat{N}$ for each peak of the mixture model for all values of p in the frequency bands 200 Hz to 400 Hz. For each p, we have either two or three estimates N_k corresponding to the number of mixture components. We can observe that from $p = 1.3$ upwards, we have a good match with the model since the middle line aligns with $N_2 = 8$. Moreover, near $p = 0.4$, the lowest estimate of N aligns with $N_1 = 8$. This indicates that there is one component in the signal which is robustly modeled by a generalized Gaussian with p values in the range 1.3 to 2.5 and another which fits the model near $p = 0.4$.

To better visualize the components present at each frequency, Fig. 5 illustrates which p-values have peaks where the estimate falls in the range $6 < \hat{N} < 10$. That is, we identify those peaks in the mixture, which correspond to components

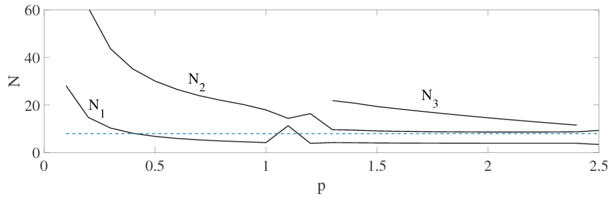

Figure 4: *Illustration of the estimated $\hat{N}$ values for each component of the mixture in the frequency band 200 Hz to 400 Hz, over all p-values.*

which follow the generalized Gaussian distribution. Each match is illustrated with a black square in the figure. Strikingly, above 4 kHz, the model fits for all $p \geq 1.3$. Since we already determined that at this range, we have only a single component in the excitation, we can conclude that the speech excitation above 4 kHz is roughly Gaussian.

In the middle range, 1 kHz to 4 kHz, we find two matches, one in the range $p = 0.5$ to 1.3, and another above $p > 1.8$, which coincides with the finding from Fig. 3, where we concluded that in this range we have two components. This would support the hypothesis that the excitation consists of noisy and harmonic components, corresponding to generalized Gaussian models with approximately $p = 2$ and $p = 0.6$. The same conclusions hold also for the lowest frequency range, below 1 kHz, although here the estimates have more variance and the findings are thus not entirely conclusive.

For the lowest frequencies the results are inconclusive also because, in contrast to our hypothesis of two excitation components, we find here consistently three components. There are several possible explanations for this deviation: Firstly, for the lowest band, 0 Hz to 200 Hz, the harmonic component, when present, will always fall in the upper part of the band. Energy content will therefore be dominant in the upper frequencies, whereby the assumption of equal variance within a band is no-longer fulfilled. Similar effects, though to a lower degree can happen in the following band as well.

Secondly, the bandwidth of peaks in the spectral envelope can, at the lower frequencies, can be of the same order as the width of frequencies bands in the analysis. This will similarly break the assumption of equal variance in the analysis band.

Finally, the speech recordings can have inaudible DC- and low-frequency components which bias the results at the low frequencies. At the high frequencies we could have similar recording artifacts, but since such distortions would likely have a broadband character, it is unlikely that they would bias the measurement.

5. Conclusions

A classic rule of thumb in speech processing is that speech signals follow the Laplace distribution [8], which does hold true for the time-signal with some restrictions. However, already when taking into account a rudimentary speech production model, things become more interesting. Assuming the linear source-filter model, speech signals can be modeled by a mixture of two excitations filtered by an IIR-filter. The time-signal is thus convolutional in character and will definitely experience interesting correlations between samples, whereby analysis of multivariate distributions becomes cumbersome.

It is possible to decorrelate the signal with, say, the Karhunen-Loève transform (KLT), but since it is based on the covariance matrix, it already makes the assumption that the signal is Gaussian. Using time-frequency transforms suffers from this restriction to a lesser degree, and the convolution becomes multiplication in the frequency domain.

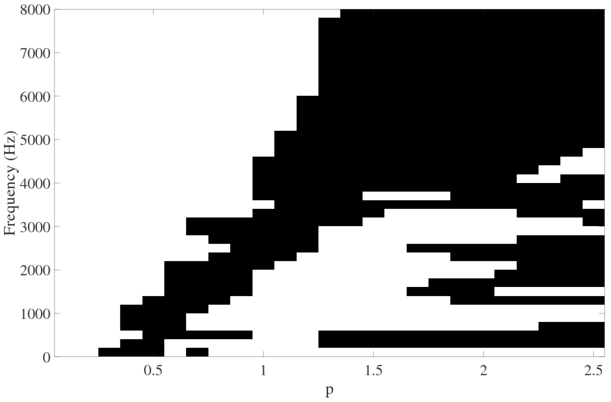

Figure 5: *Illustration of those p-values which satisfy the probability distribution model at each frequency. Specifically, if the estimated Beta mixture has a component, corresponding to a model of order $6 < \hat{N} < 10$, then the corresponding square is black.*

In the current paper we have demonstrated that we gain access to the distribution of the excitation signal by normalizing the signal spectrum in small neighborhoods (bands) of the spectrum. The analytic analysis shows that norm-ratios of all generalized Gaussian components follow Beta-distributions, whereby we can determine which distribution matches the input signal.

The presented experiments demonstrate that the excitation indeed contains at least two components. A component which is approximately Gaussian ($p > 1.8$) is present throughout the spectrum. A second component which follows the generalized Gaussian with $0.5 \leq p \leq 1.2$ is present at frequencies below 3.5 kHz. A third component, whose origin is presently unknown, is also present at the lowest frequencies (below 600 Hz).

In any case, as a rule of thumb, we can say that harmonic excitations follow approximately the Laplacian distribution and noisy excitations are Gaussian. Though this result is hardly surprising, it is more specific than prior works such as [8, 9, 10], since we include the contribution of the speech production model and can evaluate phoneme groups separately.

The distribution of the harmonic excitation, however, remains intriguing. The p value for harmonic excitation lies in the range $0.5 \leq p \leq 1.2$, which includes the value $p = 0.6$, which is known to approximate the perceptual scale of the human ear [19]. This would support the hypothesis that the human ear and speech production have been jointly tuned by evolution.

For the practical application, however, our results are comforting. The distributions of the proposed norm-ratio matches the Beta-distributions even if the value of p is chosen other than the corresponding generalized Gaussian distribution. Moreover, it is important to realize that the distribution depends on the normalization. After all, all signals are normalized to some extent, whereby the choice of normalization always affects the distribution. In applications, such as coding, we therefore do not need to know the *actual* distribution of the data, but only the distribution of the representation used.

The presented results can be used in improvement of speech processing methods which rely on statistical descriptions of the signal. For example, by employing more accurate models of the distribution, we can reduce the bit-rate of speech codecs.

6. References

[1] T. Bäckström, *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017.

[2] J. Benesty, M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer, 2008.

[3] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, 2011.

[4] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson, 2014, vol. 3.

[5] T. Bäckström and C. R. Helmrich, "Arithmetic coding of speech and audio spectra using TCX based on linear predictive spectral envelopes," in *Proc. ICASSP*, Apr. 2015, pp. 5127–5131.

[6] R. Sugiura, Y. Kamamoto, N. Harada, H. Kameoka, and T. Moriya, "Optimal coding of generalized-Gaussian-distributed frequency spectra for low-delay audio coder with powered all-pole spectrum estimation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 8, pp. 1309–1321, 2015.

[7] R. Saeidi, P. Alku, and T. Bäckström, "Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 42–53, 2016.

[8] W. B. Davenport Jr, "An experimental study of speech-wave probability distributions," *The Journal of the Acoustical Society of America*, vol. 24, no. 4, pp. 390–399, 1952.

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

[10] A. Aroudi, H. Veisi, H. Sameti, and Z. Mafakheri, "Speech signal modeling using multivariate distributions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 35, 2015.

[11] S. Korse, T. Jähnel, and T. Bäckström, "Entropy coding of spectral envelopes for speech and audio coding using distribution quantization," in *Proc. Interspeech*, 2016.

[12] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[13] C. Walck, *Handbook on statistical distributions for experimentalists*. University of Stockholm Internal Report SUF-PFY/96-01, 2007.

[14] J. S. Garofolo, L. D. Consortium *et al.*, *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, 1993.

[15] *TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)*, 3GPP, 2014.

[16] T. Bäckström, F. Ghido, and J. Fischer, "Blind recovery of perceptual models in distributed speech and audio coding," in *Proc. Interspeech*, 2016.

[17] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 2013.

[18] K. Bowman and L. Shenton, "Estimation: Method of moments," *Encyclopedia of statistical sciences*, 2004.

[19] V. Pulkki and M. Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. John Wiley & Sons, 2015.