SHAPE PARAMETER ESTIMATION FOR GENERALIZED-GAUSSIAN-DISTRIBUTED FREQUENCY SPECTRA OF AUDIO SIGNALS

Ryosuke Sugiura, Yutaka Kamamoto, Takehiro Moriya

NTT Communication Science Labs., Nippon Telegraph and Telephone Corp.,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

ABSTRACT

We have devised a method for estimating, from a single frame of audio frequency spectra, a shape parameter of multivariate generalized Gaussian distribution which has variance represented by an all-pole model and no covariance. Based on powered all-pole spectrum estimation (PAPSE), which is an extension of linear prediction, the proposed method simultaneously estimates the shape parameter and the maximum-likelihood variance, allowing more accurate representation of the probability density functions of the spectra. This paper shows an integration of the estimation into an audio codec for an example of its application, which resulted in the enhancement of the objective and subjective reconstruction quality. Since this estimation method provides us with simple parameters which reflect some acoustic features of signals, the method may also be useful in other audio signal processing problems.

***Index Terms*—** Generalized Gaussian distribution, linear prediction, feature extraction, audio compression

1. INTRODUCTION

In general, one of the main approaches for signal processing problems is to apply statistical models to the probability density functions (PDFs) of signals and estimate the parameters in these models according to the observation.

Linear prediction (LP) is a widely-used method for this approach in audio signal processing, and LP-based methods have been invented for many applications: for example, automatic speech recognition [1], audio retrieval [2], speech dereverberation [3], and audio coding [4]. In the sense of statistical modeling, LP can be interpreted as estimating the maximum-likelihood variance of multivariate Gaussian distribution (with no covariance) to which the frequency spectra of the observed signal are assumed to belong. Under this interpretation, the variance is modeled by an all-pole filter and its values also represent an envelope of the observed spectra.

Indeed the assumption of Gaussian in LP makes its model and algorithm simple, but there are many cases where it shows higher likelihood to assume other distributions. Therefore, for instance, in transform coded excitation (TCX) [5–7], an audio coding scheme which compresses signals in frequency domain, the input spectra are assumed to be Laplacian-distributed when the bits are allocated even though its variance is still estimated by LP. Our previous work [8] pointed out that, in the cases like TCX, LP does not give the maximum-likelihood variance for non-Gaussian distributions and derived a simple method of optimal estimation for those cases called powered all-pole spectrum estimation (PAPSE).

PAPSE allows us to estimate the maximum-likelihood variance for generalized Gaussian distribution (GGD), which has a parameter to control its shape (hereinafter we call it "shape parameter") to represent variant distributions including Gaussian and Laplacian. The shape parameter of GGD is related to the sparseness of the observation [3], and its applications have been studied not only for audio but also for image and video [9–13]. The work in [8] showed that PAPSE, with appropriate shape parameter, enables us to represent observed spectra with higher likelihood compared to LP in the same order, i.e., in the same degree of freedom. Moreover, the appropriate shape parameter changed momentarily depending on some kind of sparseness of the spectra and was expected to reflect the acoustic features of the observation. However, the smart method to choose the appropriate shape parameter from the observation is still unclear since it requires simultaneous estimation, from a single frame, of the shape parameter and the variance represented by PAPSE for GGD of this shape.

In this paper, LP and PAPSE are first breifly reviewed in the context of maximum-likelihood estimation. Then, we introduce a simple method for approximately estimating both the maximum-likelihood shape parameter and variance of GGD based on the PAPSE scheme. Additionally, we show an example of its application to a TCX-based codec.

2. REPRESENTING VARIANCE OF MULTIVARIATE DISTRIBUTION BY SPECTRAL ENVELOPE

2.1. Linear prediction

LP models the PDF of observed frequency spectra $\{X_k\}_{k=0}^{N-1}$ by Gaussian with its variance represented by all-pole spectra, in other words, spectral envelope $\{H_k\}_{k=0}^{N-1}$:

$$f_G(|X_k| \,||\, H_k) = \frac{1}{\sqrt{2\pi}H_k} \exp\left(-\frac{1}{2}\left|\frac{X_k}{H_k}\right|^2\right), \tag{1}$$

$$H_k^2 = \sigma^2 \left|1 + \sum_{n=1}^{p} a_n e^{-\mathrm{j}\frac{\pi nk}{N}}\right|^{-2}, \tag{2}$$

where k, N, $\{a_n\}_{n=1}^{p}$ and σ^2 respectively indicate frequency bin, frame length, LP coefficients and the power of prediction error. The maximum-likelihood estimates of the model parameters, i.e., LP coefficients $\{a_n\}_{n=1}^{p}$ are given by solving the following problem:

$$\min_{\{a_n\}} \sum_{k=0}^{N-1} D_{\mathrm{IS}}(H_k^2 \,||\, |X_k|^2) \tag{3}$$

where $D_{\mathrm{IS}}(x||y) = y/x - \ln(y/x) - 1$ is called Itakura-Saito (IS) divergence. This minimization problem is known to be, for the sake of the all-pole properties, equivalent to minimizing prediction error in the time domain and can be solved efficiently by Levinson-Durbin algorithm, which is reviewed in [8].

2.2. Powered all-pole spectrum estimation

PAPSE was intended to represent the maximum-likelihood estimate of variance for GGD with a given shape parameter α [8]:

$$f_{GG}(|X_k| \,||\, H_{k,\alpha}, \alpha) = \frac{A(\alpha)}{H_{k,\alpha}} \exp\left(-\left|B(\alpha)\frac{X_k}{H_{k,\alpha}}\right|^{\alpha}\right) \quad (4)$$

where A, B are constants written by gamma function $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\mathrm{d}t$ as

$$A(\alpha) = \frac{\alpha B(\alpha)}{2\Gamma(1/\alpha)}, \;\; B(\alpha) = \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}. \quad (5)$$

Slightly modifying the representation of the spectral envelope from LP as

$$H_{k,\alpha}^2 \equiv \left(\frac{\alpha B(\alpha)^\alpha \sigma^2}{|1+\sum_{n=1}^p a_n e^{-\mathrm{j}\frac{\pi nk}{N}}|^2}\right)^{\frac{2}{\alpha}} = (\alpha^{\frac{1}{\alpha}} B(\alpha) H_k^{\frac{2}{\alpha}})^2, \quad (6)$$

the minimization problem for estimating the maximum-likelihood $\{a_n\}_{n=1}^p$ becomes as

$$\min_{\{a_n\}} \sum_{k=0}^{N-1} D_{\mathrm{IS}}(H_k^2 \,||\, |X_k|^\alpha), \quad (7)$$

which is written in IS divergence and can be solved just as LP using Levinson-Durbin algorithm regarding α-th-powered spectra as power spectra.

3. SHAPE PARAMETER ESTIMATION WITH PAPSE

As stated above, we can obtain, for a given shape parameter α, the maximum-likelihood variance of GGD. Here, our concern is how to find smartly, among various shape parameters in the PAPSE scheme, the best parameter to represent the observation for each frame, namely the frame-by-frame maximum-likelihood α.

Although there are some previous works on estimating the shape parameter, simply applying them to the PAPSE scheme leads to inaccurate results: Methods such as moment-based method and maximum-likelihood estimation in [14] assume uniform variance over the observation, which conflicts with the PAPSE model; Methods for multivariate GGD as in [15] require several observations belonging to the same distribution, which is hard to collect for audio signals since their distributions are varying momentarily.

Therefore, we present here an iterative algorithm for simultaneously estimating the shape parameter and the variance of GGD based on the method in [14]. This algorithm is composed of two steps: PAPSE step and shape parameter estimation step. At first, we set an initial value for α and iteratively perform the following steps for the observed spectra $\{X_k\}_{k=0}^{N-1}$:

1. (PAPSE step) Estimate the maximum-likelihood variance $\{H_{k,\alpha}\}_{k=0}^{N-1}$ by PAPSE of the present α;
2. (Shape parameter estimation step) Since the normalized spectra $\{Y_k \equiv X_k/H_{k,\alpha}\}_{k=0}^{N-1}$ has approximately uniform variance, estimate the shape parameter from $\{Y_k\}_{k=0}^{N-1}$ by the method in [14] and update α.

Our preliminary test showed that the maximum-likelihood estimation in [14] often results in negative αs, which causes computational instability. Thus, we hereinafter use the moment-based method to approximate maximum-likelihood estimate of α in the shape parameter estimation step. In principle, the moment-based method estimates α by solving

$$F(\alpha) \equiv \frac{\Gamma(2/\alpha)}{\sqrt{\Gamma(1/\alpha)\Gamma(3/\alpha)}} = \frac{m_1}{\sqrt{m_2}} \quad (8)$$

where m_1 and m_2 are respectively the empirical first and second moments of $\{Y_k\}_{k=0}^{N-1}$. Actually, it is much easier to choose α from its candidates $\{\alpha_i\}_{i=1}^I$ which makes $F(\alpha_i)$ closest to $m_1/\sqrt{m_2}$ since we cannot explicitly calculate the inverse of $F(\alpha)$.

If the method used in the shape parameter estimation step gives maximum-likelihood estimate of α, the algorithm proposed above makes the likelihood monotonically increase by the iteration, which proves its convergence. However, because of the approximation of shape parameter estimation by the moment-based method, we cannot prove the convergence or optimality of this algorithm.

4. APPLICATION TO AUDIO CODEC

One of the examples for applications of the proposed estimation method is TCX, a high-compression audio coding scheme whose fundamental framework is adopted in newly-established 3GPP enhanced voice services (EVS) standard [16–18]. TCX first represents the input signals into real-valued spectra by modified discrete cosine transform (MDCT) and then quantizes the spectra by scalar quantizer. Secondly, TCX allocates bits to quantized spectra by range-coder-based arithmetic coding [4] according to their log-likelihood, and therefore higher likelihood enables more efficient compression. The step size of the quantization is decided by a bisection search in order to meet the target bit rate so that efficient compression tends to attain high sound quality in reconstructed signals.

The calculation of the log-likelihood in [6] is done by using Laplacian distribution with its variance given by LP, of which performance was improved by using GGD of $\alpha = 0.7$ with its variance given by PAPSE [8]. Here, we can apply the idea of the shape parameter estimation: estimating frame-by-frame α which makes higher likelihood, using different distributions in the arithmetic coding according to the estimated α, and transmitting α to the decoder by allocating bits to it.

5. EVALUATION OF ESTIMATION METHOD

5.1. Likelihood comparison

At first, to check the characteristics of the proposed method, its effect on likelihood, initial dependence, convergence, required number of iteration, we estimated shape parameters for some speech and audio signals and calculated their likelihood. For the shape parameter estimation step in the proposed method, we prepared for the α candidates $\alpha = 0.1$ to $\alpha = 3$ in increments of 0.1.

Figure 1 plots the average log-likelihood. The likelihood was calculated as relative log-likelihood in bits compared by its counterpart of LP:

$$L = \frac{1}{NM}\sum_{k,l} \log_2 \frac{f_{GG}(|X_{k,l}| \,||\, H_{k,\alpha_l^*,l})}{f_G(|X_{k,l}| \,||\, H_{k,l})} \quad (9)$$

where l, M, and α_l^* respectively stands for the frame number, the total frames, and the estimated α for the l-th frame. $H_{k,\alpha_l^*,l}$ and $H_{k,l}$ were given by 16-th-order PAPSE and 16-th-order LP, respectively. It can be seen that there is some convergence and it depends on the

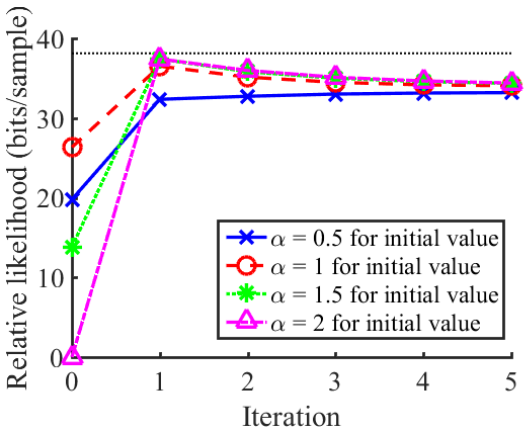

Fig. 1. Relative average log-likelihood compared to LP (bits/sample) by each iteration and initial values. Black dotted horizontal line shows the limit found by an exhaustive search. 17232 frames of 16-kHz audio signals (about 6 minutes) were tested.

Table 1. Relative average log-likelihood compared to LP (bits/sample). 17232 frames of 16-kHz audio signals (about 6 minutes) were tested.

Without PAPSE	With PAPSE		
Simple moment-based estimation	Fixed ($\alpha = 0.7$)	Proposed estimation	Exhaustive search
-60.1111	27.9153	37.4118	38.1298

initial values. Practically, it seems sufficient to use a large value for the initial α and the results of the first iteration.

Table 1 compares other estimation methods with the proposed method (the result of the first iteration with initial value $\alpha = 2$). The difference between the proposed method and the exhaustive search was less than 1 bit, and changing the shape parameter frame by frame showed higher likelihood compared to the fixed PAPSE, which calculated the log-likelihood with a constant α for every frames. Simple moment-based method, which estimated α using the moment-based method without normalizing the spectra by its variance $H_{k,\alpha}$, seems to have given inaccurate estimates resulting in the decrease of the likelihood from LP.

5.2. Transition of shape parameters

Next we compared the transition of the estimated shape parameters. The test signal was composed of four seconds each of a popular music (synthesizer), a classic music (violin), a jazz music (trumpet), a male speech (clean), and the same male speech with pink noise (signal-to-noise ratio was 10 dB), which was used in [8]. Figure 2 depicts the result. The shape parameters estimated by the proposed method roughly tracked the optimal shape parameters found by the exhaustive search, showing the correspondence between the shape parameter and some acoustic features. On the other hand, the simple moment-based method constantly estimated as around $\alpha = 0.5$, which revealed to be inaccurate by the previous experiment.

6. EVALUATION OF AUDIO CODEC

6.1. Codec settings

To evaluate the proposed method in the audio codec, we prepared a TCX-based codec which is the same one reviewed in [8]. At 16.4 kbps, we used two settings for the comparison:

1. (Baseline TCX) Using GGD of $\alpha = 0.7$ for arithmetic coding with its variance represented by fixed PAPSE of 16-th order, of which coefficients $\{a_n\}_{n=1}^{16}$ were vector quantized in the form of LSP as in [19] with 20 bits;
2. (Adaptive PAPSE) Adding shape parameter estimation part before PAPSE. The arithmetic coding used different GGD for each frame based on the estimated α. The variance was also represented by 16-th-order PAPSE, of which coefficients were quantized by the method stated above with different codebooks for each α. The α was represented by 1 bit as mentioned below.

This baseline TCX is the low-delay audio codec which was created in [8] and showed comparable subjective quality to 3GPP extended adaptive multirate wide-band (AMR-WB+) [20] at same bit rate with significantly lower coding delay. For the estimated shape parameter, we used the results from the first iteration of the proposed method with initial value $\alpha = 1$. The quantization of the estimated shape parameter α was designed heuristically by trial and error, with the optimal α expected to change smoothly in audio signals: Representing the quantized shape parameter in τ-th frame $\hat{\alpha_\tau}$ with fourth-order moving average by 1 bit, in other words, selecting $\hat{\beta}_\tau$ which satisfies

$$\hat{\alpha_\tau} = \mu^{-1}(\hat{\beta}_\tau + 0.7\hat{\beta}_{\tau-1} + 0.6\hat{\beta}_{\tau-2} + 0.5\hat{\beta}_{\tau-3} + 0.4\hat{\beta}_{\tau-4}) \quad (10)$$

where $\mu(\alpha)$ indicates the μ-law algorithm of ITU-T G. 711 [21]. The values for $\hat{\beta}_\tau$ was defined to make $\hat{\alpha_\tau}$ be in $[0.5\ \ 1]$.

The other bit allocations were evenly set: 8 bits for the step size, 3 bits for the noise-fill level, and rest of the bits for arithmetic coding (see [8, 22] for details).

6.2. Objective evaluation

The objective sound quality of the reconstructed signals, graded from -4 to 0 points, was calculated by McGill University's AFsp PQevalAudio [23]. Since TCX is mainly expected to compress audio signals, we evaluated the sound quality using musical data: Fifty items randomly selected from the four databases in the RWC Music Database [24]: Ten items each from the Classical Music, Jazz Music and Music Genre Databases; Twenty items from the Popular Music Database, ten without vocals and ten with vocals. Ten seconds of signals were extracted from each items and down-sampled into 16 kHz.

The relative scores of the TCX with adaptive PAPSE compared to the baseline TCX are shown in Fig. 3. The shape parameter estimation made the average objective quality higher, giving a significant difference in the total score. As for the complexity, the additional computational costs for applying the shape parameter estimation in this condition were about 0.6 weighted million operations per second (WMOPS [25]), about 2 % of the total costs of the TCX-based coder.

6.3. Subjective evaluation

To evaluate whether the difference in objective quality shown by the previous experiment is actually audible, an informal subjective evaluation was held. Five audio items in the RWC Music Database, ten

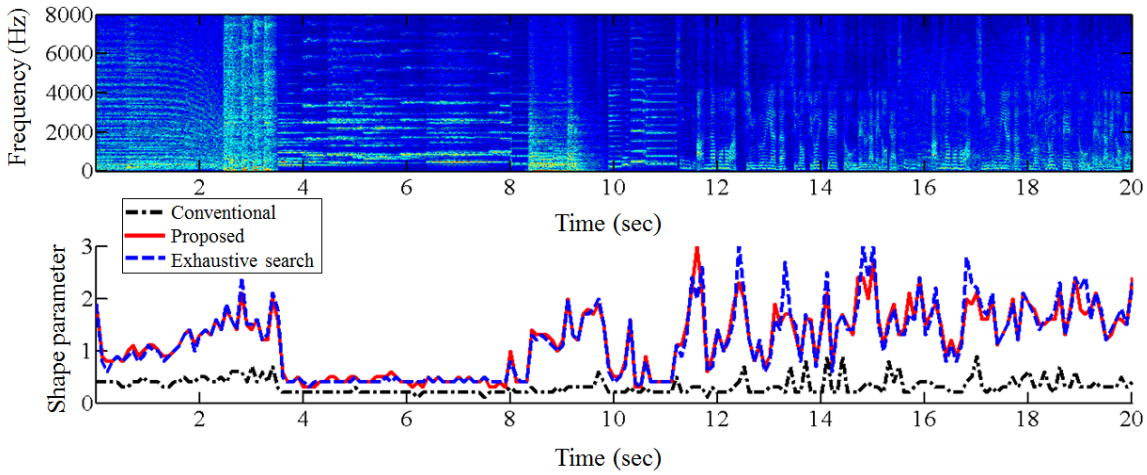

Fig. 2. Spectrogram and its shape parameters estimated for every five frames. Simple moment-based method for conventional method (black chained line), proposed method (red solid line), and exhaustive search (blue slashed line).

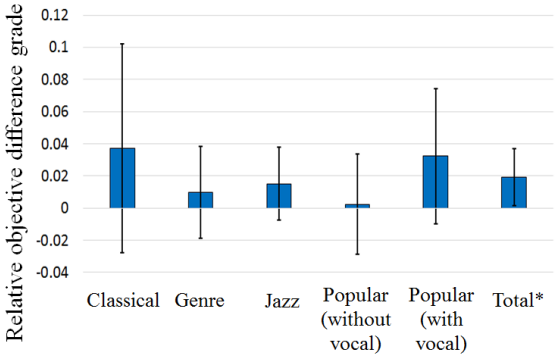

Fig. 3. Database-wise relative objective difference grades by PEAQ compared to the baseline TCX. Average and 95 % confidence intervals. Asterisk indicates there was a significant difference at 5 % in a paired t-test.

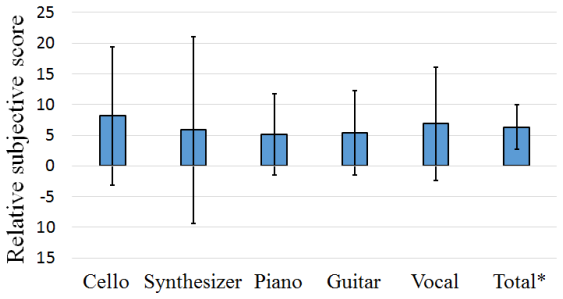

Fig. 4. Item-wise relative subjective scores by MUSHRA compared to the baseline TCX. Average and 95 % confidence intervals. Asterisk indicates there was a significant difference at 5 % in a paired t-test.

seconds each down-sampled into 16 kHz, were respectively coded in the two conditions, presented to seven participants with the references and 3.5-kHz band-limited anchors, and graded from 0 to 100 points, as is done in ITU-R BS.1534-1 Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) [26]. The test items were labeled as follows: ”Cello”, for a cello piece from the Classical Music Database, ”Synthesizer”, for a synthesizer piece from the Music Genre Database, ”Piano”, for a piano piece from the Jazz Music Database, ”Guitar”, for a Guitar piece from the Popular Music Database, and ”Vocal”, for a female vocal piece from the Popular Music Database.

Figure 4 describes the item-wise relative scores of the TCX with adaptive PAPSE compared to the baseline TCX. It can be seen that the average subjective quality were enhanced by the shape parameter estimation with a significant difference in the total score, which resembles the results of the objective evaluation.

7. CONCLUSION

We presented a simple way to estimate both the shape parameter and variance of generalized Gaussian distribution using powered all-pole spectrum estimation (PAPSE) and the moment-based shape parameter estimation. This estimation scheme enables us to represent more precisely, in the sense of likelihood, the distributions of audio frequency spectra, which have nonuniform variance over the frequencies. Despite that the proposed estimation is just an approximation for maximum-likelihood estimation, the results of the experiments proved that the estimates gave near log-likelihood to the optimal ones found by the exhaustive search and that they actually enhanced the objective and subjective quality of an audio coder. The parameters brought by this estimation may be useful in other tasks like automatic speech recognition, which would be a future challenge.

8. REFERENCES

[1] S. Soni, "Speech recognition by linear prediction," *International Journal of Scientific and Engineering Research*, vol. 6, pp. 210–216, May 2015.

[2] P.S. Jadhav and S.H. Deshmukh, "A survey on audio retrieval system for classification," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, pp. 6808–6812, Nov. 2014.

[3] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[4] S. Salomon and G. Motta, *Handbook of Data Compression*, chapter 10, Springer, 2010.

[5] G. Fuchs, C.R. Helmrich, G. Markovic, M. Neusinger, E. Ravelli, and T. Moriya, "Low delay LPC and MDCT-based audio coding in the EVS codec," in *Proc. ICASSP 2015*, Apr. 2015, pp. 5723–5727.

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

[7] "3GPP TS 26.445 release 12," 3GPP, 2014.

[8] 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," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 8, pp. 1309–1321, Aug. 2015.

[9] R. Yu, X. Lin, S. Rahardja, and C.C. Ko, "A statistics study of the mdct coefficient distribution for audio," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, Jun. 2004, vol. 2, pp. 1483–1486.

[10] M. Oger, S. Ragot, and M. Antonini, "Transform audio coding with arithmetic-coded scalar quantization and model-based bit allocation," in *Proc. ICASSP 2007*, Apr. 2007, vol. 4, pp. IV–545–IV–548.

[11] C. Bouman and K. Sauer, "A generalized gaussian image model for edge-preserving map estimation," *Image Processing, IEEE Transactions on*, vol. 2, no. 3, pp. 296–310, Jul. 1993.

[12] P. Moulin and Juan L., "Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors," *Information Theory, IEEE Transactions on*, vol. 45, no. 3, pp. 909–919, Apr. 1999.

[13] C. Parisot, M. Antonini, and M. Barlaud, "3D scan-based wavelet transform and quality control for video coding," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 56–65, 2003.

[14] S. Yu, A. Zhang, and H. Li, "A review of estimating the shape parameter of generalized gaussian distribution," *Journal of Computational Information Systems*, vol. 8, no. 21, pp. 9055–9064, 2012.

[15] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, "Parameter estimation for multivariate generalized gaussian distributions," *Signal Processing, IEEE Transactions on*, vol. 61, no. 23, pp. 5960–5971, Dec. 2013.

[16] "3GPP TS 26.441 release 12," 3GPP, 2014.

[17] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, Lei Miao, Zhe Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, Hosang Sung, Eunmi Oh, Hao Yuan, and Changbao Zhu, "Overview of the EVS codec architecture," in *Proc. ICASSP 2015*, Apr. 2015, pp. 5698–5702.

[18] S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Jarvinen, L. Laaksonen, N. Harada, N. Naka, S. Ragot, S. Proust, T. Sanda, I. Varga, C. Greer, M. Jelinek, M. Xie, and P. Usai, "Standardization of the new 3GPP EVS codec," in *Proc. ICASSP 2015*, Apr. 2015, pp. 5703–5707.

[19] H. Ohmuro, T. Moriya, K. Mano, and S. Miki, "Vector quantization of LSP parameters using moving average interframe prediction," *Electronics and Communications in Japan*, vol. 77, no. 10, pp. 12–26, 1994.

[20] "3GPP TS 26.290 release 11," 3GPP, 2012.

[21] ITU-T Rec. G. 711, "Pulse code modulation (PCM) of voice frequencies," 1988.

[22] R. Sugiura, Y. Kamamoto, N. Harada, H. Kameoka, and T. Moriya, "Resolution warped spectral representation for low-delay and low-bit-rate audio coder," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 2, pp. 288–299, 2015.

[23] "AFsp PQevalAudio," [Online]. Available: `http://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/AFsp.html` (as of Aug. '16).

[24] M. Goto, "Development of the RWC music database," in *the 18th International Congress on Acoustics (ICA 2004)*, 2004, vol. 1, pp. 553–556.

[25] "ITU-T Software Tool Library 2009 User's Manual," 2009.

[26] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2001.