# Empirical Distributions of DFT-Domain Speech Coefficients Based on Estimated Speech Variances

Timo Gerkmann and Rainer Martin
Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, 44780 Bochum, Germany
Email: {timo.gerkmann,rainer.martin}@rub.de

***Abstract*—We present a novel way to estimate the empirical distribution of clean speech spectral coefficients. Rather than computing the histogram of clean speech within a certain signal-to-noise ratio interval, we normalize the spectral coefficients on the square-root of the spectral variance estimated via recursive averaging, the decision-directed approach or temporal cepstrum smoothing. We show that estimated distributions depend significantly on the used spectral variance estimator. Further, if the speech spectral variance is estimated in noisy conditions, the resulting histograms exhibit heavier tails as compared to clean conditions. The cepstral variance estimation approach is shown to result in less heavy tails as compared to the decision-directed approach.**

## I. Introduction

In this work we discuss the estimation of the distribution of speech spectral coefficients obtained from a short-time discrete Fourier transform (DFT) in the context of speech enhancement frameworks. The distribution of clean speech spectral coefficients is of great importance for speech enhancement algorithms. It is used to derive an optimal estimator for the clean speech spectral coefficients given a noisy observation. If both clean speech and noise spectral coefficients are Gaussian distributed, the optimal estimator for the clean speech spectral coefficients is the well-known Wiener filter. Other estimators result if an estimate of the clean speech spectral amplitudes or their logarithm is desired [1], [2]. More recently, also various estimators for supergaussian distributed speech spectral coefficients have been derived [3], [4], [5], [6], [7]. Due to the nonstationary character of speech, the determination of the true local distribution is very difficult. To obtain an estimate for the distribution, the histogram of the spectral coefficients is used. However, it is important that the samples used to create the histogram obey the same distribution model and, in particular, have the same mean and variance. While it is reasonable to assume that complex speech spectral coefficients have zero mean, the determination of the variance is particularly difficult.

To determine the variance of speech spectral coefficients, the ensemble average of the magnitude squared spectral coefficients has to be taken [8]. For speech signals, however, only a single realization is available at each time-frequency point. Therefore, the ensemble average has to be replaced by a smoothing over time and/or frequency. As speech is nonstationary and hence not ergodic, the variance estimation can only be an approximation of the true variance. As a consequence, also the estimated distribution of speech spectral coefficients will have errors. While due to the central limit theorem and the sum inherent in the DFT, for increasing segment sizes the speech spectral coefficients are asymptotically complex Gaussian distributed, in speech processing frameworks segment sizes are usually too small for the central limit theorem to hold. This was observed by Porter and Boll [9] and confirmed in [3], [4].

In this work we review the estimation of the distribution of clean speech spectral coefficients used in [3], [4] in Section II, while in Section III we propose an alternative method.

## II. Review on estimating and modeling the distribution of clean speech

In this section, we review the method to estimate the histogram of clean speech proposed in [3], [4], and also review a parameterized distribution function that can be fitted to observed histograms.

The distribution of clean speech is used as prior knowledge when deriving optimal estimators for clean speech given a noisy observation. The prior should reflect the distribution of clean speech when a specific estimator is used for speech variance estimation. In single channel speech enhancement often the *decision-directed* approach [1] is used for speech variance estimation. When the histogram method is used to describe the distribution of a random process, great care has to be taken that the samples used for creating the histogram obey the same distribution model. Thus, Martin [3] proceeds as follows: white noise is added to a clean speech signal at 40 dB global signal-to-noise ratio (SNR) and the *a priori* SNR is estimated via the decision-directed approach. Then, to assure that the samples used for creating the histogram have a similar variance, the histograms are evaluated only in a narrow interval where the estimated *a priori* SNR is between 28 and 30 dB. There are some limitations using this procedure. First, the resulting histograms depend on the chosen SNR range as reported in [7]. Secondly, the interval cannot be chosen infinitely small, which means that the resulting histogram will be taken over random processes with different variances, such that the histogram might appear to be more super-Gaussian than it should. If for example the realizations of two Gaussian processes with a variance of 28 dB and 30 dB are concatenated, the resulting kurtosis is $\mathrm{kurt} \approx 3.16$ and thus larger than for each of the Gaussian processes ($\mathrm{kurt} = 3$), indicating a super-Gaussian distribution. Finally, evaluating the histogram only in an interval of 28 to 30 dB of the estimated SNR is rather restrictive: while low energy speech components that yield an *a priori* SNR below 28 dB do not contribute to the histograms at all, due to the one-frame delay of the decision-directed SNR estimate, even silence may contribute at speech offsets, where the decision-directed *a priori* SNR estimate may still be large while the speech sound is not active anymore. On the other hand, speech onsets do not fully contribute to the histograms, as at speech onsets the decision-directed SNR estimate is still rather low. The fact that silence is included in the histograms will result in an overestimation of the frequency of small amplitudes.

To derive optimal clean speech estimators, the spectral amplitudes can be modelled by the generalized Gamma distribution

$$p(|S_k|) = \frac{\nu}{\Gamma(\mu)} \left( \frac{\mu}{\sigma^2_{\mathrm{S},k}} \right)^{\mu} |S_k|^{\nu\mu-1} \exp\left( -\frac{\mu}{\sigma^2_{\mathrm{S},k}} |S_k|^{\nu} \right) , \qquad (1)$$

where, $S_k$ is the clean speech spectral coefficient at frequency bin $k$, $\sigma^2_{\mathrm{S},k} = \mathrm{E}\left\{ |S_k|^2 \right\}$ is the variance of the complex spectral coefficients, $\mu$ and $\nu$ are shape parameters, and $\Gamma(\cdot)$ is the complete gamma function [10, (8.31)]. For $\nu = 2$ (1) resembles the $\chi$-distribution, while for $\nu = 1$ (1) resembles the $\chi^2$-distribution. In this context $2\mu$ is often referred to as the degrees of freedom [11]. For complex Gaussian distributed spectral coefficients $S_k$, the spectral amplitudes are distributed as (1) with $\mu = 1$, $\nu = 2$, which is also referred to as Rayleigh distribution. Super-Gaussian distributions can be modelled by choosing $\nu < 2$ or $\mu < 1$. Optimal estimators for $\chi$ and $\chi^2$-distributed spectral amplitudes are given by Erkelens, Hendriks *et al.* [6], Andrianakis and White [7], and Breithaupt *et al.* [12].

## III. Proposed estimation of the distribution of clean speech

In this section we propose an alternative way to estimating the distribution of speech spectral coefficients.

A minimum mean square error (MMSE) optimal estimate for a function of clean speech spectral coefficients is given as [9], [11], [13]

$$\mathrm{E}\Big\{c(S_k(l))|Y_k(l)\,,\sigma^2_{\mathrm{S},k}(l),\sigma^2_{\mathrm{N},k}(l),\mathcal{H}_{1,k}(l)\Big\}\,, \tag{2}$$

where $Y_k$ is the noisy observation, $l$ is the segment index, $\sigma^2_{\mathrm{N},k}$ is the noise variance, $\mathcal{H}_{1,k}$ is the hypothesis that speech is present, and $c(\cdot)$ is some function, such as the absolute value operator, the squared absolute value, or the logarithm of the absolute value. To solve the expectation in (2) the distribution of the clean speech spectral coefficients is needed. When the distribution of clean speech spectral coefficients is determined in the context of MMSE estimation, two important properties can be seen from (2). First, (2) is conditioned on speech presence. Thus, the histogram should be computed using all (but only) those time-frequency points where speech is active. Secondly, (2) is conditioned on the speech spectral variance. Thus, the speech spectral variance is assumed to be *known* at each time-frequency point, and is thus treated as deterministic. Considering the first property, we propose to include all time-frequency points where the magnitude squares of clean speech spectral coefficients are larger than -65 dB with respect to the largest time-frequency coefficient in a considered speech sample, similar to [6]. This is in contrast to [3] where the histogram considers only few speech (and possibly also non-speech) spectral coefficients where the estimated variance is large. To consider the deterministic character of the speech spectral variance in (2), we propose to normalize the complex spectral coefficients on the square-root of an estimate of the speech variance. This is different to [3] where the histogram is computed over a certain interval of the *a priori* SNR, and thus a certain variation of the speech spectral variance can be expected. The histogram obtained by the proposed approach thus reflects the average distribution of all clean speech spectral coefficients for a given variance estimator under the condition that speech is present, and hence allows to solve (2).

### A. Estimating the speech spectral variance

A simple variance estimator is given by a recursive averaging with a fixed smoothing factor $\alpha$, as

$$\sigma^2_{\mathrm{S},k}(l) = \alpha\,\sigma^2_{\mathrm{S},k}(l-1) + (1-\alpha)\,|S_k(l)\,|^2\,. \tag{3}$$

As the recursive estimate (3) is a function of the current clean speech coefficient $S_k(l)$, the recursive variance estimate $\sigma^2_{\mathrm{S},k}(l)$ and the current speech coefficient $S_k(l)$ are correlated. As a consequence, $|S_k(l)\,|$ cannot be arbitrarily large with respect to a given $\sigma_{\mathrm{S},k}$ estimated by (3). With respect to $|S_k(l)\,|^2$ the smallest possible value for $\sigma^2_{\mathrm{S},k}(l)$ is given when the previous estimate is zero, *i.e.* $\sigma^2_{\mathrm{S},k}(l-1) = 0$. Then, $\sigma^2_{\mathrm{S},\min,k}(l) = (1-\alpha)\,|S_k(l)\,|^2$ and the upper bound for the normalized amplitudes is

$$\frac{|S_k(l)\,|}{\sigma_{\mathrm{S},\min,k}(l)} = \frac{1}{\sqrt{1-\alpha}}. \tag{4}$$

As a consequence, no heavy tails can be expected when the recursive averaging is used to estimate the speech spectral variance.

To reduce a smearing of the speech spectral structure, for speech enhancement algorithms, often the decision-directed approach [1] is used to estimate the variance of the speech spectral coefficients. With the observed noisy speech $Y_k(l) = S_k(l) + N_k(l)$ and the noise variance $\sigma^2_{\mathrm{N},k}$, the decision-directed speech spectral variance estimator can be written as

$$\begin{aligned}\sigma^2_{\mathrm{S},k}(l) = \alpha_{\mathrm{dd}}\,|G_k(l-1)\,Y_k(l-1)\,|^2 \\ + (1-\alpha_{\mathrm{dd}})\,\max\{0, |Y_k(l)\,|^2 - \sigma^2_{\mathrm{N},k}(l)\}.\end{aligned} \tag{5}$$

In the case the noise signal is zero, we have $Y_k = S_k$, $\sigma^2_{\mathrm{N},k} = 0$, and $G_k = 1$. For clean speech, the decision-directed speech variance estimator thus results in

$$\sigma^2_{\mathrm{S},k}(l) = \alpha_{\mathrm{dd}}\,|S_k(l-1)\,|^2 + (1-\alpha_{\mathrm{dd}})\,|S_k(l)\,|^2 \tag{6}$$

Further, as usually $\alpha_{\mathrm{dd}}$ is chosen close to one, for clean speech, the decision-directed approach results approximately in the periodogram of the previous frame, as

$$\sigma^2_{\mathrm{S},k}(l) \approx |S_k(l-1)\,|^2. \tag{7}$$

Note that the decision-directed approach can also be defined without this delay [14, Section 3.4.1]. Then, given a clean observation, the histogram of the spectral amplitudes would be approximately given by a single peak at $|S_k(l)\,|/\sigma_{\mathrm{S},k} = 1$.

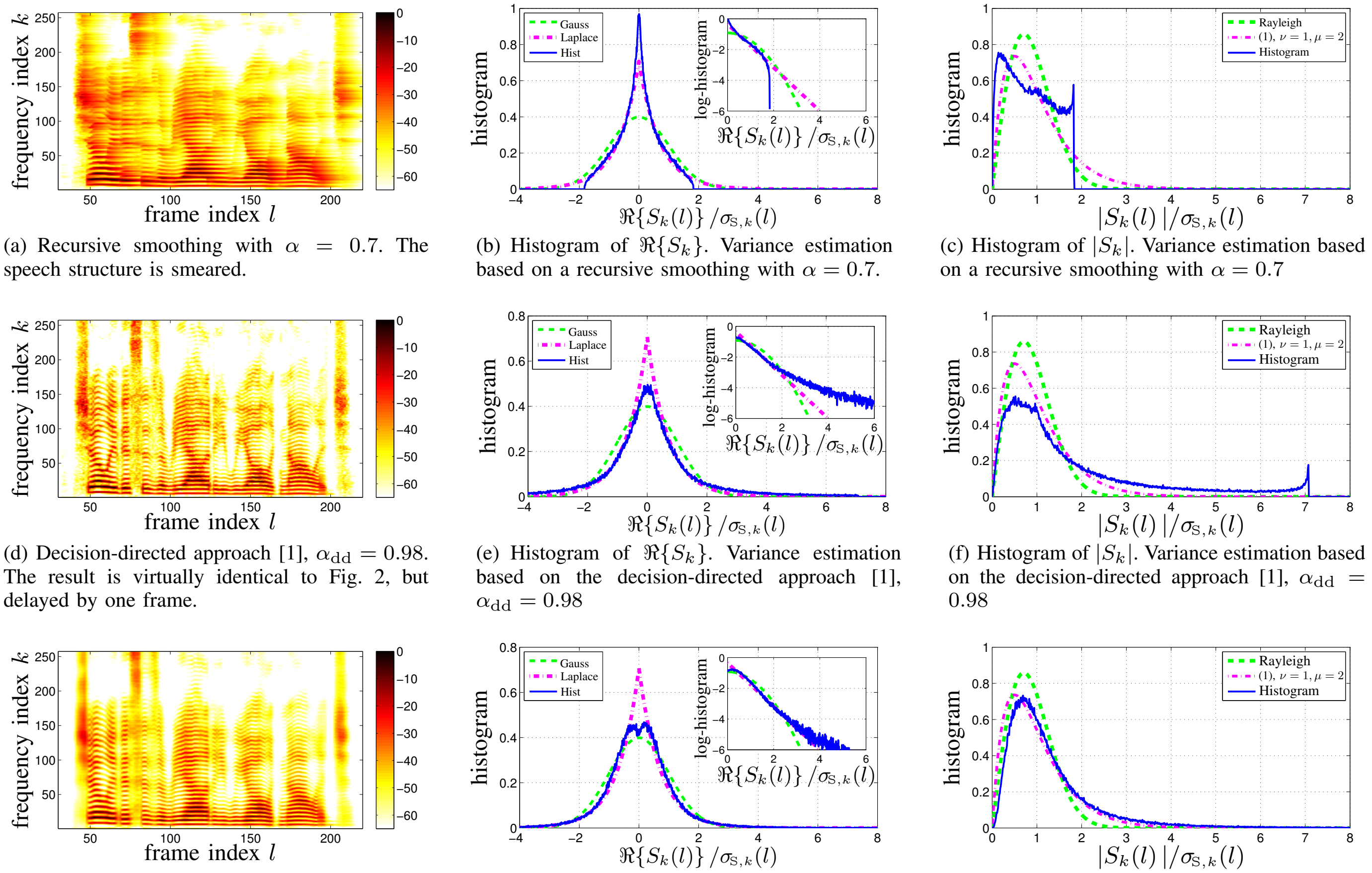
The third variance estimator we consider is the temporal cepstrum smoothing approach proposed in [15]. For the bias compensation we use [16, Algorithm 1] and assume that the speech spectral amplitudes are $\chi$-distributed with $\mu = 1$.

### B. Observed distribution given a clean observation

In Fig. 1 the results of the three variance estimators for the clean speech in Fig. 2 are given. For creating the histograms, we use 5 male and 5 female speakers from the TIMIT database [17], a sampling rate of $f_\mathrm{s} = 16$ kHz, a segment length of 32 ms, and a Hann spectral analysis window without zero-padding and with 50% overlap. We show the Gaussian and the super-Gaussian Laplace distribution with unit variance along with the histogram of the real part of the speech spectral coefficients. The histogram of the spectral amplitudes is shown along with a Rayleigh-distribution and unit variance, and the $\chi^2$-distribution with shape parameter $\mu = 2$ and unit variance.

It can be seen that the recursive smoothing smears speech onsets over time (compare Fig. 2 and Fig. 1(a)). The resulting histogram for the recursive variance estimator is very peaked, but does not exhibit heavy tails, as the normalized amplitudes are bound to be smaller than $1/\sqrt{1-\alpha} = 1.823$. The decision-directed approach yields basically a shifted version of the input (compare Fig. 2 and Fig. 1(d)). The histogram of the real part of the complex spectral coefficients in Fig. 1(e) is more peaked than the Gaussian distribution but less peaked than the Laplace distribution. Further, the histogram is clearly more heavy-tailed than the Gaussian and the Laplace distribution. Accordingly, also the histogram of the spectral amplitudes in Fig. 1(c) exhibits heavy tails as compared to the Rayleigh distribution.

The cepstral approach in Fig. 1(g) can be seen to result in a smoothing of $|S_k|^2$ that maintains the speech spectral structure very nicely. The tails of the histogram of the real part in Fig. 1(h) can be well modeled by the Laplace distribution, while the tails of the spectral amplitudes in Fig. 1(i) are well modeled by a $\chi^2$-distribution and $\mu = 2$. Using the cepstral approach can be seen to yield less heavy tails as compared to the histograms obtained by the decision-directed approach.



(a) Recursive smoothing with $\alpha = 0.7$. The speech structure is smeared.

(b) Histogram of $\Re\{S_k\}$. Variance estimation based on a recursive smoothing with $\alpha = 0.7$.

(c) Histogram of $|S_k|$. Variance estimation based on a recursive smoothing with $\alpha = 0.7$

(d) Decision-directed approach [1], $\alpha_{\mathrm{dd}} = 0.98$. The result is virtually identical to Fig. 2, but delayed by one frame.

(e) Histogram of $\Re\{S_k\}$. Variance estimation based on the decision-directed approach [1], $\alpha_{\mathrm{dd}} = 0.98$

(f) Histogram of $|S_k|$. Variance estimation based on the decision-directed approach [1], $\alpha_{\mathrm{dd}} = 0.98$

(g) Temporal cepstrum smoothing [15]. The speech spectral structure is well preserved.

(h) Histogram of $\Re\{S_k\}$. Variance estimation based on temporal cepstrum smoothing [15].

(i) Histogram of $|S_k|$. Variance estimation based on temporal cepstrum smoothing [15].

Fig. 1. Spectrograms and histograms for different speech variance estimators. The clean speech is given in Fig. 2. While in the first line a recursive smoothing with $\alpha = 0.7$ is used, the second line gives the results for the decision-directed approach. In the last line the result for temporal cepstrum smoothing are shown. In the left column the speech variance estimate is plotted, in the middle column the histogram of the real part of clean speech is compared to a Gaussian and a Laplace distribution. In the right column the histogram of the amplitude of clean speech is compared to a Rayleigh distribution and a $\chi^2$-distribution with shape parameter $\mu = 2$.

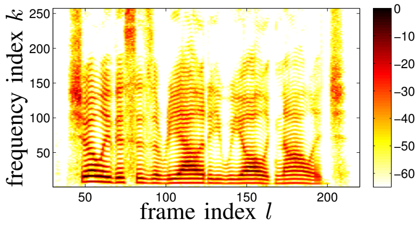
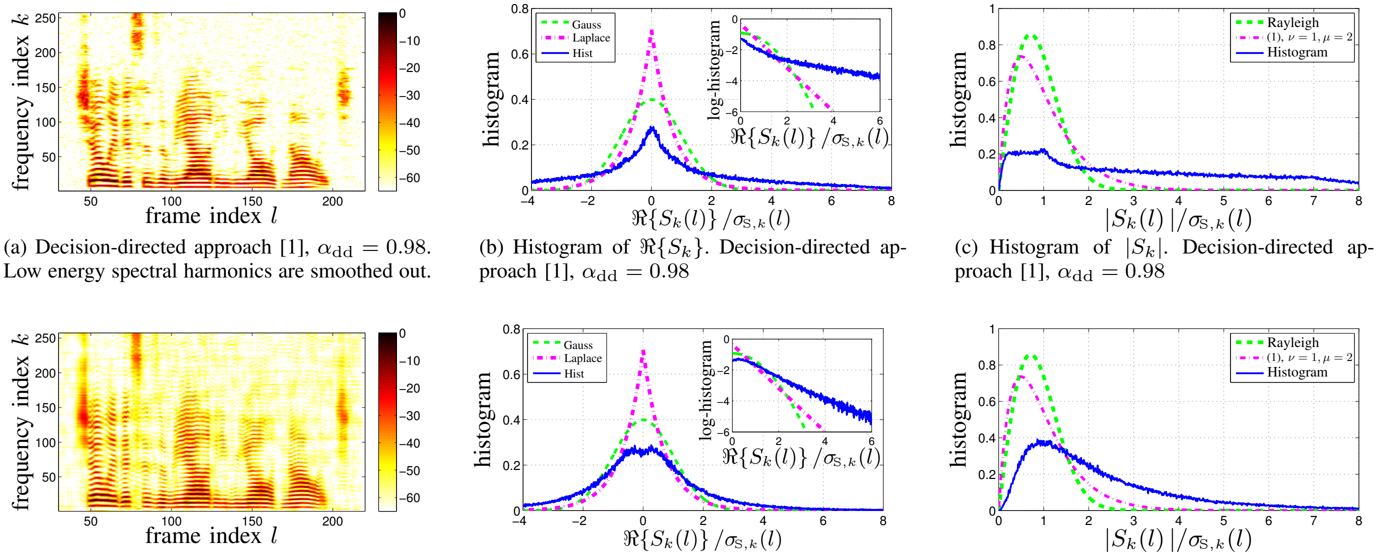


Fig. 2. Clean speech spectral coefficients $|S_k(l)|^2$.

### C. Observed distribution given a noisy observation

In the previous section we estimated the histograms of clean speech, based on the speech variance $\sigma^2_{S,k}$ which was estimated given clean speech. We have seen that the shape of the distribution clearly depends on the chosen speech variance estimator. However, in a noisy environment, the speech variance estimate on which (2) is conditioned, is estimated from noisy speech. Thus, one may argue that the distribution of clean speech, which is used to solve (2), should be a distribution that results from the speech spectral variance estimated in a noisy environment. To see how the histograms change when the speech power is estimated in a noisy environment, we degrade clean speech by white Gaussian noise at a segmental SNR of 10 dB. For the decision-directed approach (5) we employ the Wiener filter $G_k(l-1) = \sigma^2_{S,k}(l-1)/(\sigma^2_{S,k}(l-1)+\sigma^2_{N,k}(l-1))$. While the resulting speech variance estimates and histograms are given in Fig. 3, the noisy speech is given in Fig. 4.

Comparing the estimated speech variance in the presence of noise in Fig. 3 to the estimated speech variance in the absence of noise in Fig. 1, it can be seen that at time-frequency points where speech is present, the speech spectral variance is often underestimated if it is estimated in a noisy environment. The underestimation of the spectral variance results in a larger likelihood that speech spectral coefficients are larger than the square-root of the spectral variance, which in turn results in heavier tails of the histograms (compare Fig. 3 to Fig. 1). Comparing Fig. 2 to Fig. 3(a) and Fig. 3(d), it may be seen that the cepstral approach preserves the speech spectral structure better than the decision-directed approach. Thus, also in the noisy case the cepstral approach yields less heavy tails as compared to the decision-directed approach.



(a) Decision-directed approach [1], $\alpha_{\mathrm{dd}} = 0.98$. Low energy spectral harmonics are smoothed out.

(b) Histogram of $\Re\{S_k\}$. Decision-directed approach [1], $\alpha_{\mathrm{dd}} = 0.98$

(c) Histogram of $|S_k|$. Decision-directed approach [1], $\alpha_{\mathrm{dd}} = 0.98$

(d) Temporal cepstrum smoothing [15]. The speech spectral structure is well preserved.

(e) Histogram of $\Re\{S_k\}$. Variance estimation based on temporal cepstrum smoothing [15].

(f) Histogram of $|S_k|$. Variance estimation based on temporal cepstrum smoothing [15].

Fig. 3. Similar setup as in Fig. 1 but the speech spectral variance is estimated speech disturbed by white Gaussian noise at 10 dB input SNR (cf. Fig. 4).

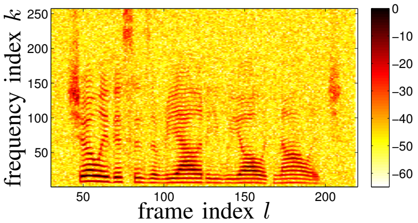


Fig. 4. Noisy speech spectral coefficients $|Y_k(l)|^2$. The speech coefficients are disturbed by white Gaussian noise at 0 dB segmental SNR.

## IV. Conclusion

We reconsidered the estimation of the distribution of clean speech spectral coefficients used for a minimum mean square error (MMSE) estimation of clean speech spectral coefficients. As the considered MMSE estimators treat the speech spectral variance as deterministic, the speech variance should also be treated as deterministic when estimating the underlying distribution. For this we propose to compute the histograms of all clean speech spectral coefficients normalized on the square-root of the estimated speech spectral variance. We have shown that the empirical distribution depends significantly on the used variance estimator. Further, as in the presence of noise state-of-the-art speech variance estimators tend to underestimate the speech variance in speech presence, results indicate that with a decreasing global SNR more heavy-tailed speech priors should be used.

## References

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

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Process*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," *IEEE ICASSP*, pp. 253–256, May 2002.

[4] ——, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[5] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Applied Signal Process*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.

[6] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[7] I. Andrianakis and P. R. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *ELSEVIER Speech Communication*, vol. 51, no. 1, pp. 1–14, Jan. 2009.

[8] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York, NY, USA: McGraw-Hill, 2002.

[9] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," *IEEE ICASSP*, pp. 18A.2.1–18A.2.4, 1984.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals Series and Products*, 6th ed. San Diego, CA, USA: Academic Press, 2000.

[11] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. Chichester, West Sussex, UK: John Wiley & Sons, 2006.

[12] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," *IEEE ICASSP*, pp. 4037–4040, Apr. 2008.

[13] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Heidelberg, Germany: Springer Verlag, 2008, ch. 44, pp. 873–901.

[14] C. Breithaupt, "Noise reduction algorithms for speech communications – statistical analysis and improved estimation procedures," Ph.D. dissertation, Ruhr-Universität Bochum, Bochum, Germany, 2008.

[15] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," *IEEE ICASSP*, pp. 4897–4900, Apr. 2008.

[16] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[17] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, 1988.