Spectral Domain Speech Enhancement using HMM State-Dependent Super-Gaussian Priors

Nasser Mohammadiha, *Student Member*, IEEE, Rainer Martin, *Fellow*, IEEE, and Arne Leijon, *Member*, IEEE

Abstract

The derivation of MMSE estimators for the DFT coefficients of speech signals, given an observed noisy signal and super-Gaussian prior distributions, has received a lot of interest recently. In this letter, we look at the distribution of the periodogram coefficients of different phonemes, and show that they have a gamma distribution with shape parameters less than one. This verifies that the DFT coefficients for not only the whole speech signal but also for individual phonemes have super-Gaussian distributions. We develop a spectral domain speech enhancement algorithm, and derive hidden Markov model (HMM) based MMSE estimators for speech periodogram coefficients under this gamma assumption in both a high uniform resolution and a reduced-resolution Mel domain. The simulations show that the performance is improved using a gamma distribution compared to the exponential case. Moreover, we show that, even though beneficial in some aspects, the Mel-domain processing does not lead to better results than the algorithms in the high-resolution domain.

Index Terms

HMM, super-Gaussian pdf, speech enhancement.

I. Introduction

Time-frequency domain single-channel noise reduction approaches using super-Gaussian priors have received a lot of attention during recent years. The real and imaginary parts of the speech (and noise)

N. Mohammadiha and A. Leijon are with the Sound and Image Processing Laboratory, School of Electrical Engineering, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: nmoh@kth.se; leijon@kth.se).

R. Martin is with the Institute of Communication Acoustics (IKA) Ruhr-University Bochum, 44780 Bochum, Germany (e-mail: rainer.martin@rub.de).

DFT coefficients, for instance, are better modeled with super-Gaussian distributions, e.g. Laplacian and two-sided gamma distributions, than with a Gaussian distribution [1]. Several approaches have been proposed to derive MMSE estimators for the DFT coefficients of speech, given the noisy signal, using these super-Gaussian prior distributions [1]–[3]. In these works the super-Gaussianity is considered for the long-term statistics of a speech signal and not conditioned on the phoneme type. Hence, an interesting question is whether this phenomenon depends on the phoneme type.

Moreover, it is important for signal processing algorithms to investigate the distribution of the speech and noise DFT coefficients given the so called "hidden state", which can be considered as the phoneme type. This can be very beneficial in deriving better estimators in the HMM-based speech enhancement approaches [4]–[7]. Traditionally, HMM-based noise reduction schemes have been derived by assuming auto-regressive (AR) models for the speech and noise signals, and then the AR parameters are assumed to be Gaussian [4]–[6]. Recently, an HMM-based speech enhancement approach was proposed in [8] in which the DFT coefficients of the speech and noise signals were assumed to be complex Gaussian.

This letter proposes two main contributions:

1) We explore the distribution of the state-conditional speech DFT coefficients. Our experiments show that phoneme-dependent periodogram coefficients have a gamma (with shape parameters less than one) rather than an exponential distribution.

2) We extend the HMM-based speech enhancement algorithm from [8] and derive new MMSE estimators for the speech power spectral coefficients using super-Gaussian prior distributions, given the noisy signal. We assume that the speech power spectral coefficients are gamma-distributed while noise power spectral coefficients are Erlang-distributed. Our simulations show that the performance of the proposed denoising algorithm is superior to algorithms using the exponential distribution. Hence, the results support the super-Gaussianity hypothesis. Furthermore, we compare the performance of the derived estimators in the high-resolution DFT domain and in the reduced-resolution Mel frequency domain.

II. Conditional Distribution of the Speech Power Spectral Coefficients

In this section, we look at the distribution of the speech power spectral coefficients – estimated using periodogram or magnitude-squared DFT coefficients – conditioned on the hidden state that can be seen as the phoneme type. We denote the random variables associated with the speech DFT coefficients and their realizations by $\bar{O}_{mt}$ and $\bar{o}_{mt}$, respectively, where m is the frequency bin and t is the time-frame index. Moreover, let $|\cdot|^2$ represent the element-wise magnitude-square operator. Let us define the conditional

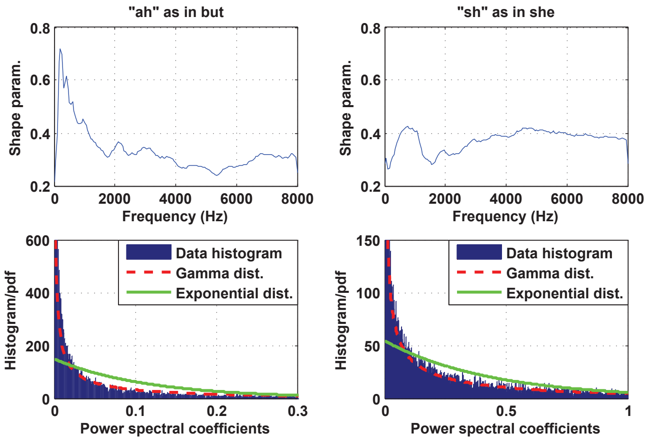

Fig. 1. Experimental distribution of the speech power spectral coefficients for two speech sounds “ah” and “sh”. The top panel shows the estimated shape parameters of the fitted gamma distributions at different frequencies. The bottom panel shows the histogram and estimated distributions of spectral coefficients of ”ah” at 2500 Hz (left), and of ”sh” at 6000 Hz (right).

gamma distribution as:

$$f\left(|\bar{o}_{mt}|^2 \mid \bar{S}_t = i\right) = \frac{\left(|\bar{o}_{mt}|^2\right)^{\alpha_{mi}-1}}{\left(b_{mi}\right)^{\alpha_{mi}}\,\Gamma\left(\alpha_{mi}\right)} e^{-|\bar{o}_{mt}|^2/b_{mi}}, \tag{1}$$

where $\bar{S}_t = \bar{s}_t \in [1, \bar{N}]$ is the hidden state, α_{mi} and b_{mi} are the state-dependent shape and scale parameters, and $\Gamma(\cdot)$ is the complete Gamma function. If $\bar{N} = 50 \sim 60$, each state is identified roughly by one phoneme. For (1), we have: $E(|\bar{O}_{mt}|^2 \mid \bar{S}_t = i) = \alpha_{mi} b_{mi}$ and $\mathrm{var}(|\bar{O}_{mt}|^2 \mid \bar{S}_t = i) = \alpha_{mi} b_{mi}^2$.

For $\alpha_{mi} = 1$, (1) reduces to an exponential distribution. This corresponds to assuming that real and imaginary parts of the DFT coefficients ($\mathrm{Re}\{\bar{O}_{mt}\}$ and $\mathrm{Im}\{\bar{O}_{mt}\}$) have a Gaussian distribution. For $\alpha_{mi} < 1$, however, the resulting distribution for DFT coefficients will be super-Gaussian, as shown next. Assuming that $\mathrm{Re}\{\bar{O}_{mt}\}$ and $\mathrm{Im}\{\bar{O}_{mt}\}$ are independent and identically distributed, Eq. (1) leads to a gamma distribution for $|\mathrm{Re}\{\bar{O}_{mt}\}|^2$ and $|\mathrm{Im}\{\bar{O}_{mt}\}|^2$ with shape parameters equal to $\alpha_{mi}/2$. This is because the sum of two independent gamma random variables (RV) with equal scales is a gamma RV. Then, it can be easily shown that $|\mathrm{Re}\{\bar{O}_{mt}\}|$ and $|\mathrm{Im}\{\bar{O}_{mt}\}|$ have generalized gamma distribution with $\nu = \alpha_{mi}/2, \gamma = 2, \beta = 1/b_{mi}$ (see [2] for definition of these parameters), or equivalently, $\mathrm{Re}\{\bar{O}_{mt}\}$ and $\mathrm{Im}\{\bar{O}_{mt}\}$ have two-sided generalized gamma distributions.

A. *Experimental Data*

To obtain the experimental phoneme-conditioned distribution of the speech power spectral coefficients, we used 2000 realizations for each phoneme from the TIMIT database at a sampling rate of 16 kHz. The waveform of each realization was normalized to have unit variance. To obtain the spectral coefficients, first, each waveform was windowed into short-time frames using a Hann window with a frame length of 20 ms and 50% overlap, and second, the DFT was applied to these short-time frames to obtain the periodogram.

The top panel of Fig. 1 shows the shape parameters of the estimated gamma distributions for two phonemes, “ah” and “sh”. The estimation of the shape and scale parameters of the gamma distributions was done using a standard maximum-likelihood approach independently for each frequency bin. As Fig. 1 shows, the shape parameters for these two phonemes are less than one at all frequencies. In the bottom panel of Fig. 1, the histogram of the power spectral coefficients of “ah” at frequency 2500 Hz (left) and of “sh” at frequency 6000 Hz (right) are shown. Also, the estimated gamma and exponential distributions are shown in Fig. 1 for comparison. As a result, we find that the power spectral coefficients will have gamma rather than exponential distributions even if we limit the speech data to come from a specific phoneme and normalize each realization. Therefore, real and imaginary parts of the phoneme-conditioned speech DFT coefficients have super-Gaussian distributions. As the top-left panel of Fig. 1 shows, there are distinct differences between phones: the shape parameters of gamma distributions corresponding to “ah” are higher at frequencies close to the main formants due to less variation in the signal energy in these frequencies. This can be generalized to other vowels as well.

III. HMM-based Speech Enhancement

A. Speech Model

Besides DFT coefficients, we also consider a more coarse resolution frequency as it reduces the number of model parameters and smoothes out the signals’ random variations before processing. In this case, the power at adjacent speech DFT bins is summed to obtain $\mathbf{X} = [X_{kt}]$, with elements representing the frame band power in analysis band k, as

$$X_{kt} = \sum_{m=m_l(k)}^{m_h(k)} w_m(k) \left|\bar{O}_{mt}\right|^2, \tag{2}$$

where w denotes a set of overlapped triangular filters that approximate the Mel-scale filter bank, m_l and m_h represent the band-dependent lowest and highest DFT indices to be summed, respectively. If $m_l = m_h$ and $w_m = 1$, we recover the original spectra as: $X_{kt} = \left|\bar{O}_{kt}\right|^2$. The state dependent conditional distribution of X_{kt} is now obtained by a slight modification of (1):

$$f\big(x_{kt} \,|\, \bar{S}_t = i, G_t = g_t\big) = \frac{(x_{kt})^{\alpha_{ki}-1}}{(g_t b_{ki})^{\alpha_{ki}}\,\Gamma(\alpha_{ki})} e^{-x_{kt}/(g_t b_{ki})}, \tag{3}$$

where G_t is a short-term stochastic gain parameter, which is assumed to have a gamma distribution as:

$$f(g_t) = \frac{g_t^{\phi-1}}{\theta_t^{\phi}\Gamma(\phi)} e^{-g_t/\theta_t}. \tag{4}$$

Here, ϕ is the shape parameter and θ_t is a time-varying scale parameter, which models the long-term speech level. Assuming conditional independence of the elements of $\mathbf{X}_t$ [1], [2], the HMM output density functions for a given state can be expressed as:

$$f(\mathbf{x}_t \mid \bar{S}_t = i, G_t = g_t) = \prod_{k=1}^{K} f\left(x_{kt} \mid \bar{S}_t = i, G_t = g_t\right). \tag{5}$$

The sequence of the speech hidden states are characterized by a fully connected first-order Markov model with transition probability matrix $\bar{\mathbf{a}}$, with elements $\bar{a}_{i'i} = f\left[\bar{S}_t = i \mid \bar{S}_{t-1} = i'\right]$, and a time-invariant state probability mass vector $\bar{\mathbf{p}}$, with elements $\bar{p}_i = f\left[\bar{S}_t = i\right]$. The parameters of the speech model denoted by $\lambda = \{\bar{\mathbf{a}}, \mathbf{b}, \boldsymbol{\alpha}, \phi, \theta\}$ are obtained from training data using the EM algorithm [8][1].

B. Noise Model

Let $\ddot{\mathbf{O}} = [\ddot{O}_{mt}]$ denote the noise DFT coefficients. The noise band power spectral vectors, $\mathbf{V} = [V_{kt}]$, are obtained similarly to (2). The noise signal is modeled using an $\ddot{N}$-state HMM with hidden states denoted as $\ddot{S}_t$. The noise power spectral coefficients are assumed to have an Erlang distribution which includes the exponential distribution as a special case and provides a sufficiently accurate fit to the data:

$$f\left(v_{kt} \mid \ddot{S}_t = j, H_t = h_t\right) = \frac{(v_{kt})^{\beta_k - 1} e^{-v_{kt}/(h_t c_{kj})}}{(h_t c_{kj})^{\beta_k} (\beta_k - 1)!}, \tag{6}$$

where β_k is the state-independent integer shape parameter, c_{kj} is the scale parameter, and ”!” represents the factorial function. The short-term stochastic gain parameter of the noise is also assumed to have a gamma distribution as:

$$f(h_t) = \frac{h_t^{\psi-1}}{\gamma_t^{\psi} \Gamma(\psi)} e^{-h_t/\gamma_t}. \tag{7}$$

The noise Markov chain construction and parameter estimation is done similarly to speech model (Subsection III-A). The only difference is that after each iteration of the EM algorithm, the shape parameters are rounded to the closest integer numbers.

C. Speech Estimation: Complex Gaussian Case

This subsection presents a speech enhancement algorithm in the DFT domain, i.e., a special case of (2) where $X_{mt} = |\bar{O}_{mt}|^2$ and $V_{mt} = |\ddot{O}_{mt}|^2$. Assuming that the DFT coefficients of the clean speech and noise signals are complex Gaussian ($\alpha_{ki} = \beta_k = 1$ in (3), (6)), DFT coefficients of the mixed signal $\mathbf{O}$, $\mathbf{O}_t = \bar{\mathbf{O}}_t + \ddot{\mathbf{O}}_t$, will also have complex Gaussian distributions. Let us represent the composite hidden

[1] The update equation of the speech shape parameters has to be modified slightly to exclude the summation over the states since $\boldsymbol{\alpha}$ are state-dependent.

state of the mixed signal by S_t with realizations s_t that can take one of the $\bar{N}\ddot{N}$ possible outcomes. Let $\sigma^2_{O_{mt}}{=}E(X_{mt} \mid \bar{s}_t, g_t){+}E(V_{mt} \mid \ddot{s}_t, h_t)$, which is calculated considering (3) and (6). We have:

$$f\left(o_{mt} \mid g_t, h_t, s_t\right) = \frac{1}{\pi\sigma^2_{O_{mt}}} e^{-\frac{|o_{mt}|^2}{\sigma^2_{O_{mt}}}}, \tag{8}$$

$$f\left(\mathbf{o}_t \mid g_t, h_t, s_t\right) = \prod_m f\left(o_{mt} \mid g_t, h_t, s_t\right). \tag{9}$$

To prevent the numerical problems, (9) is computed in the logarithmic domain. We approximate the state-conditional distribution of the mixed signal by taking a point estimate for the gain parameters (see [6], [8]), as:

$$f(\mathbf{o}_t \,|\, s_t){=}\iint f(\mathbf{o}_t, g_t, h_t \,|\, s_t) dg_t dh_t {\approx} f(\mathbf{o}_t \,|\, g'_t, h'_t, s_t). \tag{10}$$

In this letter, we use the mean values of the gain distributions as the point estimates, $g'_t = \phi\theta_t$, and $h'_t = \psi\gamma_t$. θ_t and γ_t represent the long-term speech and noise levels, respectively. As it is shown in [8], the MMSE estimator of the speech DFT coefficients is given by

$$E\left(\bar{\mathbf{O}}_t \mid \mathbf{o}_1^t\right) = \frac{\sum_{s_t} \boldsymbol{\zeta}_t(s_t, \mathbf{o}_t) E\left(\bar{\mathbf{O}}_t \mid \mathbf{o}_t, g'_t, h'_t, s_t\right)}{\sum_{s_t} \boldsymbol{\zeta}_t\left(s_t, \mathbf{o}_t\right)}. \tag{11}$$

where $\mathbf{o}_1^t{=}\{\mathbf{o}_1, ...\mathbf{o}_t\}$, and $\boldsymbol{\zeta}_t(s_t, \mathbf{o}_t){=}f(s_t|\mathbf{o}_1^{t-1})f(\mathbf{o}_t|g'_t, h'_t, s_t)$. Also,

$$f\left(s_t \mid \mathbf{o}_1^{t-1}\right) = \sum_{s_{t-1}} a_{s_{t-1},s_t} f\left(s_{t-1} \mid \mathbf{o}_1^{t-1}\right), \tag{12}$$

with $a_{s_{t-1},s_t} {=} \bar{a}_{\bar{s}_{t-1},\bar{s}_t} \ddot{a}_{\ddot{s}_{t-1},\ddot{s}_t}$, and $f(s_{t-1} \,|\, \mathbf{o}_1^{t-1})$ is the scaled forward variable. Due to the Gaussian assumptions, the state-conditional estimates of the speech DFT coefficients are obtained using a Wiener filter as:

$$E\big(\bar{O}_{mt} \,|\, \mathbf{o}_t, g'_t, h'_t, s_t\big) {=} \frac{E\left(X_{mt} \mid g'_t, \bar{s}_t\right) o_{mt}}{E(X_{mt} \,|\, g'_t, \bar{s}_t) {+} E(V_{mt} \,|\, h'_t, \ddot{s}_t)}, \tag{13}$$

in which (3) and (6) are used to compute the expected values. Although different functions of the speech DFT coefficients can also be estimated within the HMM framework [4], we used (13) here since it is a widely used reference method.

D. Speech Estimation: Erlang-Gamma Case

This subsection presents one of the main contributions of this letter where we derive new MMSE estimators using super-Gaussian prior distributions. We assume that the speech and noise band powers are additive, i.e. $\mathbf{Y}_t = \mathbf{X}_t + \mathbf{V}_t$, where the band powers are obtained similarly to (2), and $\mathbf{X}_t$, $\mathbf{V}_t$ are assumed to be independent. The additivity assumption is widely used in the literature to circumvent the difficulty of phase modeling. Here, we derive an MMSE estimator for $\mathbf{X}_t$ given that both $\mathbf{X}$ and $\mathbf{V}$ are modeled using an HMM with gamma and Erlang output distributions, respectively. Again, denote the composite hidden state of the mixed signal $\mathbf{Y}$ by $\mathbf{S}$ with $S_t = s_t \in [1, \bar{N}\ddot{N}]$. Although the conditional distribution $f\left(y_{kt} \mid g_t, h_t, s_t\right)$ is not exactly gamma, still a gamma distribution would be flexible enough to describe Y_{kt} practically, and we continue with this approximation for simplicity. Therefore, we follow a standard moment matching algorithm – up to second order, and considering that $E(Y_{kt} \mid g_t, h_t, s_t) = E(X_{kt} \mid g_t, \bar{s}_t) + E(V_{kt} \mid h_t, \ddot{s}_t)$, and var$(Y_{kt} \mid g_t, h_t, s_t) =$ var$(X_{kt} \mid g_t, \bar{s}_t) +$ var$(V_{kt} \mid h_t, \ddot{s}_t)$ – to obtain a gamma distribution to describe $f\left(y_{kt} \mid g_t, h_t, s_t\right)$. Then, the state-conditional distribution of the mixed signal is obtained using similar assumptions exploited in (9) and (10).

The MMSE estimate of the speech band powers can now be obtained similarly to (11). Since different speech band-powers are assumed to be conditionally independent, we now focus on obtaining $E(X_{kt} \mid y_{kt}, g'_t, h'_t, s_t)$. First, note that

$$f(y_{kt} \,|\, x_{kt}, h_t, s_t) = \begin{cases} f(V_{kt} = y_{kt} - x_{kt} \,|\, h_t, \ddot{s}_t) & y_{kt} \geq x_{kt}, \\ 0 & y_{kt} < x_{kt}. \end{cases} \tag{14}$$

Using Bayes rule, the MMSE estimate of X_{kt} is given as:

$$\hat{x}_{kt} = E(X_{kt} \mid y_{kt}, g'_t, h'_t, s_t) = \frac{\int_0^{y_{kt}} x_{kt} f\left(y_{kt} \mid x_{kt}, h'_t, s_t\right) f\left(x_{kt} \mid \bar{s}_t, g'_t\right) dx_{kt}}{\int_0^{y_{kt}} f\left(y_{kt} \mid x_{kt}, h'_t, s_t\right) f\left(x_{kt} \mid \bar{s}_t, g'_t\right) dx_{kt}}. \tag{15}$$

Exploiting (3),(6), and (14) in (15) yields:

$$\hat{x}_{kt} = \frac{\int_0^{y_{kt}} x_{kt}^{\alpha_{ki}} (y_{kt} - x_{kt})^{\beta_k - 1} e^{-\left(\frac{y_{kt} - x_{kt}}{h'_t c_{kj}} + \frac{x_{kt}}{g'_t b_{ki}}\right)} dx_{kt}}{\int_0^{y_{kt}} x_{kt}^{\alpha_{ki} - 1} (y_{kt} - x_{kt})^{\beta_k - 1} e^{-\left(\frac{y_{kt} - x_{kt}}{h'_t c_{kj}} + \frac{x_{kt}}{g'_t b_{ki}}\right)} dx_{kt}}, \tag{16}$$

where we have set $\bar{s}_t = i$ and $\ddot{s}_t = j$ to keep notations uncluttered. Since V_{kt} is assumed to have an Erlang distribution, β_k is integer. Using the binomial theorem, we can write:

$$(y_{kt} - x_{kt})^{\beta_k - 1} = \sum_{l=0}^{\beta_k - 1} \binom{\beta_k - 1}{l} y_{kt}^{\beta_k - 1 - l} \left(-x_{kt}\right)^l, \tag{17}$$

in which $\binom{\beta_k - 1}{l}$ is the binomial coefficient. Define $z_{k,ij} = 1/(g'_t b_{ki}) - 1/(h'_t c_{kj})$ and $\mathfrak{a}_{kl} = (-1)^l \binom{\beta_k - 1}{l} y_{kt}^{\beta_k - 1 - l}$. Since the integration and summation are interchangeable, inserting (17) into (16) yields:

$$\hat{x}_{kt} = \frac{\sum_{l=0}^{\beta_k - 1} \mathfrak{a}_{kl} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l} e^{-z_{k,ij} x_{kt}} dx_{kt}}{\sum_{l=0}^{\beta_k - 1} \mathfrak{a}_{kl} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l - 1} e^{-z_{k,ij} x_{kt}} dx_{kt}}. \tag{18}$$

In the following, we discuss two special cases for which the integrals in (18) can be solved analytically. First, for positive $z_{k,ij}$, we obtain the subsequent closed-form expression:

$$\begin{aligned} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l} e^{-z_{k,ij} x_{kt}} dx_{kt} &= z_{k,ij}^{-(\alpha_{ki} + l + 1)} \int_0^{z_{k,ij} y_{kt}} u^{\alpha_{ki} + l} e^{-u} du \\ &= z_{k,ij}^{-(\alpha_{ki} + l + 1)} \gamma\left(\alpha_{ki} + l + 1, z_{k,ij} y_{kt}\right), \end{aligned} \tag{19}$$

where we have defined $u = z_{k,ij}x_{kt}$ and $\gamma(a, y)$ is the incomplete gamma function [9, eq. 8.350].

Second, when the speech shape parameters α_{ki} are integer-valued, we use [9, eq. 2.323] to get the following closed-form solution for the required integrations in (18):

$$\int_0^{y_{kt}} x_{kt}^{\alpha_{ki}+l} e^{-z_{k,ij}x_{kt}} dx_{kt} = \frac{e^{-z_{k,ij}y_{kt}}}{-z_{k,ij}} \sum_{q=0}^{\alpha_{ki}+l} (-1)^q \frac{P^{(q)}(y_{kt})}{(-z_{k,ij})^q} + \frac{1}{z_{k,ij}} (z_{k,ij})^{-l-\alpha_{ki}} P^{(\alpha_{ki}+l)}(0), \tag{20}$$

where $P^q(y_{kt})$ is the q^{th} derivative of $x_{kt}^{\alpha_{ki}+l}$ with respect to x_{kt}, evaluated at y_{kt}.

If neither (19) nor (20) can be used to calculate (18), the integrals can be tabulated, or they can be computed online using the stochastic integrations.

The derived algorithm in this subsection provides an MMSE estimator for the speech band powers, $\hat{x}_{kt}$. To obtain an estimate of the speech spectral vectors in the original DFT resolution, we first obtain the gain function at the central frequencies of the bands as $\kappa_{kt} = \hat{x}_{kt}/y_{kt}$, and then interpolate this gain values to obtain the high resolution gain vector $\bar{\kappa}_{mt}$, and then speech DFT coefficients are estimated as $\bar{\kappa}_{mt}o_{mt}$.

IV. Experiments and Results

The proposed speech enhancement strategies are evaluated and compared at different input signal to noise ratios (*SNR*) for different interfering noise types including babble, factory and highway traffic noises. The speech models are trained using the training data from the TIMIT database while babble and factory noises were taken from NOISEX-92, and highway traffic noise was taken from Sound-Ideas database. All of the signals were down-sampled to 16-kHz. The core test set of the TIMIT database (192 sentences) was exploited for the noise reduction evaluation, and the train and test segments of noises were disjoint. The signal synthesis was performed using the overlap-and-add procedure using a frame length of 320 samples with 50% overlapped windows and a Hann window. For the speech model $\bar{N} = 55$ states and for each noise type $\ddot{N} = 10$ states were trained.

Two objective measures including source to distortion ratio (*SDR*) and perceptual evaluation of speech quality (*PESQ*) were considered for the evaluation. The *SDR* and *PESQ* improvements are averaged over all of the three noise types and the final scores are shown in Fig. 2. Three algorithms are considered for comparison. Two algorithms are implemented directly in the high-resolution spectral domain, which are referred as: complex Gaussian (13) and exp-gamma (18 with $\beta_k = 1$). To evaluate (18), we used either (19) or (20) whenever possible, and if none of them were applicable, we calculated the integrals using

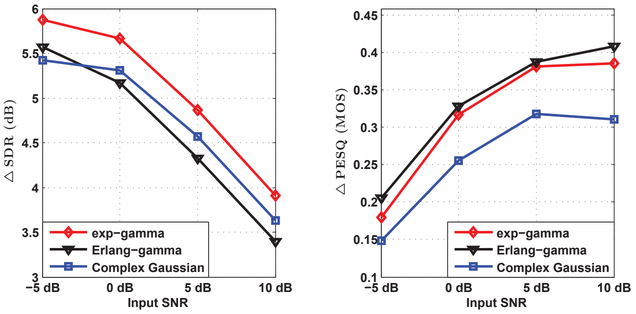

Fig. 2. Performance of the proposed noise reduction algorithms averaged over different noise types.

the stochastic integrations. The other algorithm, referred as Erlang-gamma, is implemented in reduced resolution domain, for which (18) is used.

The presented results in Fig. 2 show that the exp-gamma algorithm is clearly better than the complex Gaussian, in terms of both *SDR* and *PESQ*. Thus, the simulation results verify the observation from Section II, and imply that the real and imaginary parts of the speech DFT coefficients are modeled better with super-Gaussian than with Gaussian distributions.

The results of the Mel-domain Erlang-gamma algorithm and the DFT domain exp-gamma algorithm are very close in the sense of *PESQ*, but exp-gamma is superior considering *SDR*. The benefit of Mel-domain algorithms is that the random speech and noise fluctuations at different frequency bins are reduced and smoother signals are fed into the models. Also, the assumption of additive speech and noise band powers is more justified in this case. On the other hand, due to the reduced resolution, the filter estimation is less accurate. Informal listening test results were consistent with these objective results.

V. Conclusion

In this letter, we aim to investigate the distribution of the phoneme-conditioned speech power spectral coefficients. We looked at the empirical distribution of the periodogram coefficients for different phonemes, and also we derived new HMM-based speech spectral enhancement algorithms. The empirical distributions together with the simulation results of the denoising algorithms support our hypothesis that the power spectral coefficients will rather have gamma distributions with shape parameters less than one even at the scale of individual phones. For example, using a gamma assumption the source to distortion ratio was increased up to 0.8 dB compared to the exponential assumption. We also showed that this finding can be equivalently expressed as the super-Gaussianity of the DFT coefficients for different phonemes.

References

[1] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 13, no. 5, pp. 845–856, sep. 2005.

[2] J. S. Erkelens *et al.*, “Minimum Mean-Square Error estimation of discrete Fourier coefficients with generalized Gamma priors,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, pp. 1741–1752, 2007.

[3] B. Chen and P. C. Loizou, “A Laplacian-based MMSE estimator for speech enhancement,” *Speech Communication*, vol. 49, no. 2, pp. 134–143, feb. 2007.

[4] Y. Ephraim, “A bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, apr. 1992.

[5] H. Sameti *et al.*, “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, sep. 1998.

[6] D. Y. Zhao and W. B. Kleijn, “HMM-based gain modeling for enhancement of speech in noise,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 882–892, mar. 2007.

[7] H. Veisi and H. Sameti, “Speech enhancement using hidden Markov models in Mel-frequency domain,” *Speech Communication*, vol. 55, no. 2, pp. 205–220, feb. 2013.

[8] N. Mohammadiha and A. Leijon, “Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, to be published.

[9] I. Gradshteyn and I. Ryzhik, *Table of Integerals, Series, and Products*, 7th ed., A. Jeffrey and D. Zwillinger, Eds. Academic Press, feb. 2007.