PAPER *Special Section on Statistical Modeling for Speech Processing*

Gamma Modeling of Speech Power and Its On-Line Estimation for Statistical Speech Enhancement

Tran Huy DAT[†*a)], *Nonmember*, **Kazuya TAKEDA**[†], *Member*, *and* **Fumitada ITAKURA**[††], *Fellow*

SUMMARY This study shows the effectiveness of using gamma distribution in the speech power domain as a more general prior distribution for the model-based speech enhancement approaches. This model is a superset of the conventional Gaussian model of the complex spectrum and provides more accurate prior modeling when the optimal parameters are estimated. We develop a method to adapt the modeled distribution parameters from each actual noisy speech in a frame-by-frame manner. Next, we derive and investigate the minimum mean square error (MMSE) and maximum a posterior probability (MAP) estimations in different domains of speech spectral magnitude, generalized power and its logarithm, using the proposed gamma modeling. Finally, a comparative evaluation of the MAP and MMSE filters is conducted. As the MMSE estimations tend to more complicated using more general prior distributions, the MAP estimations are given in closed-form extractions and therefore are suitable in the implementation. The adaptive estimation of the modeled distribution parameters provides more accurate prior modeling and this is the principal merit of the proposed method and the reason for the better performance. From the experiments, the MAP estimation is recommended due to its high efficiency and low complexity. Among the MAP based systems, the estimation in log-magnitude domain is shown to be the best for the speech recognition as the estimation in power domain is superior for the noise reduction.

key words: *speech enhancement, speech recognition, gamma modeling, fourth-order moment, MMSE, MAP, spectral magnitude, power, log-spectral magnitude*

1. Introduction

1.1 Statistical Speech Enhancement

Noise reduction is a major problem of speech processing including speech recognition, hearing aid and mobile communication. Among the single-channel noise reduction methods, the statistical estimation in the spectral domain is shown to be the most effective [1]. Two directions of the model-based and data-driven approaches have been proposed in the literature. The model-based methods use a short-time learning of the joint distributions of the signal and noise spectra and then employ a statistical estimator for the clean speech spectrum [2]–[6], [9]–[11]. In contrast, the data-driven methods use available data to derive empirical codebook-dependent estimations [7], [8]. As the data-driven methods are useful for speech recognition, the model-based method is more suitable for real-time systems, when the environment is unknown or the training is unavailable. The modeling and fitting distributions of the speech and the noise spectra and the choice of the estimation criterion, resulting in the gain function, are two main issues of the model-based speech enhancement.

1.2 Conventional Methods

Conventional methods assume the zero-mean Gaussian distributions of the noise and the speech spectra, and therefore, only information on the variances (i.e., second-order statistics) is required in order to determine the distributions [2]–[4]. The signal and noise variances are updated frame by frame and it is often expressed via a priori and a posteriori signal to noise ratio (instantaneous SNR) [2], [3]. Using the Gaussian model, different gain functions are derived on the basic of different estimation criteria. The minimum mean square error (MMSE) estimation in the spectral component domain yields the classical Wiener filter. Ephraim and Malah (1984–1985) developed the MMSE estimations for speech spectral magnitude and its logarithm [2], [3], which were shown to be superior to the Wiener filter. Later, Wolfe and Godsill (2002) derived the maximum a posterior probability (MAP) estimation for speech spectral magnitude [4], which is simpler but of the same efficiency as the MMSE estimation. However, as the Gaussian model is suitable to model the noise spectrum, this model is not optimal for the speech signals. The reason is that the Gaussian model leads to the independence between magnitude and phase, which is unnatural for speech signal. Therefore, the speech spectrum is expected to be better modeled by non-Gaussian distributions.

1.3 Non-Gaussian Model Based Speech Enhancement

Some non-Gaussian models have been proposed in the literature. The super-Gaussian model and the Laplacian-gamma model of DFT coefficients were proposed by Lotter and Vary [5] and Martin [6]. The common point of these models is that, the signal variance is estimated and updated via priori SNR by the same way as in the Gaussian model-based systems, as a fixed distribution parameter set is applied for the whole signals. This parameter set is obtained from a histogram of DFT coefficients taken over a clean speech database. However, in real speech enhancement approaches, the distribution of speech spectrum is time-frequency depen-

Manuscript received June 11, 2005.

Manuscript revised September 30, 2005.

[†] The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.

[††] The author is with the Graduate School of Information Engineering, Meijo University, Nagoya-shi, 468–8502 Japan.

[*] Presently, with the Institute for Infocomm Research, Singapore.

a) E-mail: hdtran@i2r.a-star.edu.sg

DOI: 10.1093/ietisy/e89–d.3.1040

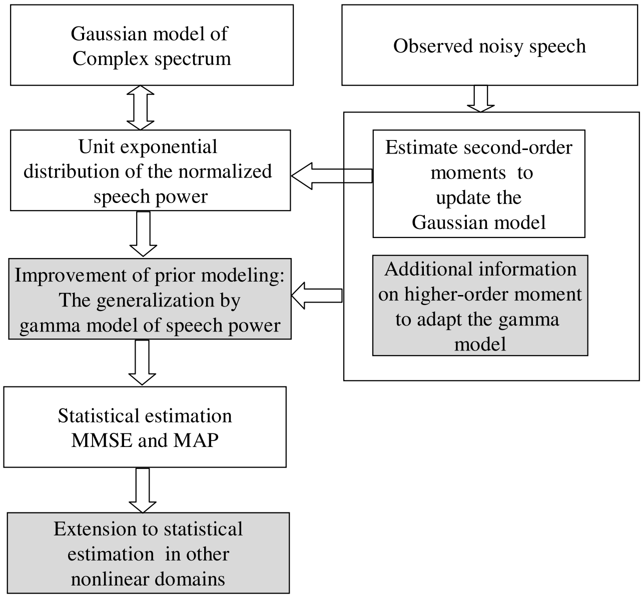
Gaussian model of Complex spectrum
Observed noisy speech
Unit exponential distribution of the normalized speech power
Estimate second-order moments to update the Gaussian model
Improvement of prior modeling: The generalization by gamma model of speech power
Additional information on higher-order moment to adapt the gamma model
Statistical estimation MMSE and MAP
Extension to statistical estimation in other nonlinear domains

Fig. 1 Statistical speech enhancement and main points of proposed system.

dent since its variance is estimated and updated time by time and therefore, the prior distribution parameters should be estimated from actual noisy speech. Moreover, the models proposed in [5] and [6] are not super-set of the conventional Gaussian model and therefore their superiority can not always be provided from the theoretical point of view, even in the case if we could estimate the parameters from noisy speech.

In this study we focus in the following issues of the non-Gaussian-model-based speech enhancement. First, we look for a more general distribution model of the speech spectrum, which should be a super-set of the conventional Gaussian model. At other-hand this model should provide a relatively low complexity in further implementation of the speech spectral estimation. We found that the gamma modeling in the speech power domain satisfies the told above requirements and this model is adopted in this study. Second, we develop on-line method to adaptively estimate the modeled distribution parameters from each actual noisy speech in a frame-by-frame manner. Third, we investigate and compare several estimation methods, performing in different domains, using the proposed prior distribution modeling. The basic idea and logical scheme-diagram of this study is summarized in Fig. 1. The organization of this paper is as follows. In Sect. 2, we address the prior distribution modeling problem, including the gamma modeling in speech power domain and its on-line parameter estimation. In Sect. 3, we derive different speech spectral magnitude estimations based on MMSE and MAP criteria in different domains using the proposed modeling. In Sect. 4, we report a comparative evaluation of the estimation methods, using the AURORA2J database. Finally, in Sect. 5, a summary of this study is presented.

2. Prior Distribution Modeling and Its On-Line Estimation

2.1 Additive Model

We consider the additive model of noisy speech in the STDFT domain

$$\mathbf{X}(n,k) = \mathbf{S}(n,k) + \mathbf{N}(n,k)\,, \tag{1}$$

where $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{N}$ are the complex spectra of noisy speech, clean speech and noise signals, respectively. Couple (n,k) denotes a frame-frequency index and will be omitted in this section. Hereafter, we use X, φ_X, S, φ_S, and N, φ_N to represent the magnitudes and phases of the complex spectra. Since the magnitude is more informative and the phase is sensitive to errors, speech enhancement systems estimate only speech spectral magnitude as the phase is remained unchanged. The key point of all algorithm is the joint distribution of noisy and clean speech magnitudes, denoted by

$$p(X,S) = p(X|S)\,p(S)\,. \tag{2}$$

Given the joint distribution, the clean speech magnitude can be estimated using a criterion such as MMSE or MAP. We will discuss this issue in the next section. From (1), the joint distribution (2) is defined by the prior distribution of speech spectral magnitude $p(S)$ and the conditional distribution $p(X|S)$, which is derived from the noise distribution. As in conventional systems [2]–[4], we assume the zero-mean Gaussian distribution of the noise spectral components

$$N_R, N_I \sim normal\left(0, \frac{\sigma_N^2}{2}\right), \tag{3}$$

where $(.)_R$, and $(.)_I$ are the real and imaginary parts of the complex spectrum. σ_N^2 is the noise spectral variance (i.e., spectral density or local power) given as

$$\sigma_N^2 = \left\langle |\mathbf{N}|^2 \right\rangle, \tag{4}$$

where $\langle\ .\ \rangle$ denotes the expectation operator. The noise variance is unknown in advance and should be estimated from noisy speech at each time-frequency index. Note that although the behaviors of the noise spectrum is various by the environments, the Gaussian model (3) is in general appropriated. The reason of this consideration is that the background noise is often presented as a superposition of a large number of random fluctuations, and following the central limit theory, its distribution should be close to the Gaussian distribution. One important property of the Gaussian model of the noise spectrum is that the joint distribution of the speech and noise spectral magnitudes is given independently from the phase component. This can be given as follows. Assumption (3) yields the joint conditional distribution of spectral components of noisy speech given the clean speech, expressed as

$$p(X_R, X_I\,|S_R, S_I) = \frac{1}{\pi\sigma_N^2}\exp\left(-\frac{{N_R}^2 + {N_I}^2}{\sigma_N^2}\right), \tag{5}$$

where $N_R = X_R - S_R$, and $N_I = X_I - S_I$. The conditional joint probability for the spectral magnitude and phase is derived using the Jacobian transform [12],

$$p(X, \varphi_X \mid S, \varphi_S) = \frac{X}{\pi\sigma_N^2} \exp\left[-\frac{X^2 - 2XS\cos(\Delta\varphi) + S^2}{\sigma_N^2}\right], \tag{6}$$

where $\Delta\varphi = \varphi_X - \varphi_N$. Integrating (6) over the noisy phase, it yields the Rician conditional probability, which is independent from noisy speech phase, noted as

$$p(X \mid S) = \frac{X}{2\pi\sigma_N^2} \exp\left(-\frac{X^2 + S^2}{\sigma_N^2}\right) I_0\left(\frac{2XS}{\sigma_N^2}\right). \tag{7}$$

Here, $I_0(x)$ is the modified Bessel function of the first-kind.

$$I_0(x) = \int_{-\pi}^{\pi} \exp\left[-x\cos(\varphi - \phi)\right] d\varphi \tag{8}$$

From (7) and (2), the joint distribution $p(X, S,)$ is independent from phase and therefore the magnitude estimation can be carried out independently. This is the key factor to consider the more preferable use of the modeling and estimation in the magnitude domain over that in the complex spectral domain as in [6].

2.2 Gamma Distribution of Speech Power

Now we turn our attention to the problem of the speech prior distribution modeling. The conventional model also assumes the zero mean Gaussian distribution of the speech spectral components,

$$S_R, S_I \sim normal\left(0, \frac{\sigma_S^2}{2}\right), \tag{9}$$

where $\sigma_S^2 = \left\langle |\mathbf{S}|^2 \right\rangle$ denotes the speech spectral variance (spectral density or local signal power), which is also estimated from actual noisy speech at each time-frequency index. As was mentioned above, the Gaussian model is not optimal for speech signal since it leads to the independence between magnitude and phase, which is unnatural for the speech signals and the aim of this study is looking for a more general model of speech prior distribution. Since the speech magnitude can be estimated independently from phase, the idea is looking for a more general distribution model in the magnitude or power domain. From the assumption (9), the Gaussian model leads to an exponential distribution of the speech power (i.e., magnitude square) and after a normalization to the local power, to a unit exponential distribution of the normalized speech power

$$p\left(\frac{S^2}{\sigma_S^2}\right) = \exp\left(-\frac{S^2}{\sigma_S^2}\right). \tag{10}$$

The distribution (10) is a special case of gamma distribution with unit parameters. From here, the gamma modeling in the power domain is a direct generalization and super-set of (10) and therefore is expected to better model the speech spectrum if the optimal parameters can be estimated. The distribution density function of gamma distribution of the normalized speech power is noted as

$$p\left(\frac{S^2}{\sigma_S^2}\right) = \frac{b^a}{\Gamma(a)} \left(\frac{S^2}{\sigma_S^2}\right)^{a-1} \exp\left(-b\frac{S^2}{\sigma_S^2}\right). \tag{11}$$

Note that (a, b) is normalized to have a unit mean

$$\left\langle S^2 \right\rangle = \sigma_S^2. \tag{12}$$

This implies a relationship between a and b expressed as

$$a = b, \tag{13}$$

and therefore, the system has only one-free parameter. Figures 2 and 3 show the distributions of the normalized speech powers, which are directly estimated from 10 dB in-car noisy speech at two frequency bins. We can see that, the gamma modeling even visually better fits the actual distributions than the unit exponential distribution derived from the conventional Gaussian model. Moreover the behaviors

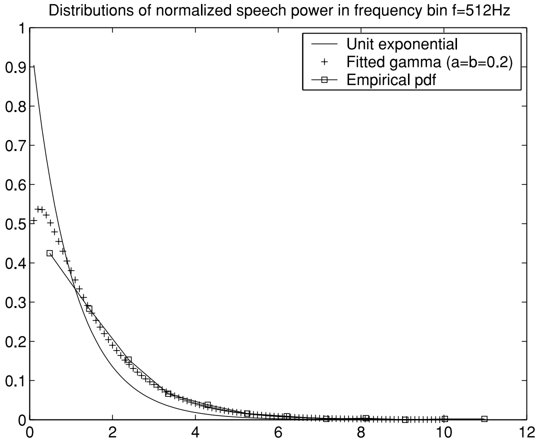

Fig. 2 Actual distribution of normalized speech power estimated from noisy speech in frequency bin $f = 512$ Hz.

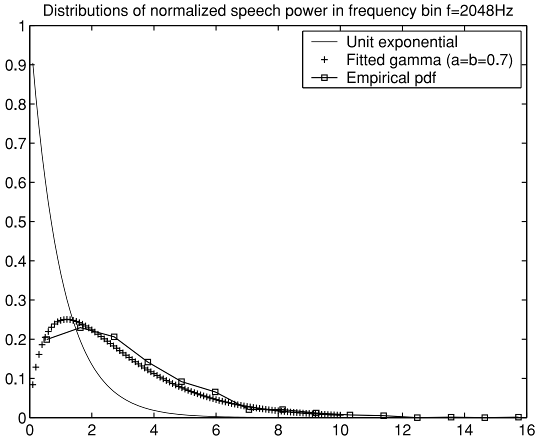

Fig. 3 Actual distribution of normalized speech power estimated from noisy speech in frequency bin $f = 2048$ Hz.

of speech powers are different in these bins.

Note that the gamma distribution (11) is equivalent to the generalized gamma distribution in the magnitude domain, denoted by

$$p\,(S) = \frac{b^a}{\Gamma\,(a)\,\sigma_S^{2a}} S^{2a-1} \exp\left(-b\frac{S^2}{\sigma_S^2}\right). \tag{14}$$

Substituting (14) and (7) into (2), we obtain the joint distribution of clean and noisy speech spectral magnitude noted as

$$p\,(X,S) = T\,(X)\,S^{2a-1}\exp\left(-\frac{S^2}{\lambda}\right) I_0\left(2S\sqrt{\frac{\vartheta}{\lambda}}\right), \tag{15}$$

where

$$T\,(X) = \frac{b^a}{\Gamma\,(a)\,\sigma_S^{2a}}\frac{X}{2\pi\sigma_N^2}\exp\left(-\frac{X^2}{\sigma_N^2}\right) \tag{16}$$

is independent of S and this term will be reduced in the MMSE and MAP estimations. Variable λ satisfies

$$\frac{1}{\lambda} = \frac{b}{\sigma_S^2} + \frac{1}{\sigma_N^2}, \tag{17}$$

and ϑ is denoted in terms of "a priori SNR" ξ and "a posteriori SNR" γ by

$$\vartheta = \frac{\xi}{\xi + b}\gamma, \quad \xi = \frac{\sigma_S^2}{\sigma_N^2}, \quad \gamma = \frac{X^2}{\sigma_N^2}. \tag{18}$$

2.3 On-Line Parameter Estimation

The main point of this study is that we consider the estimation of the modeled distribution parameters from actual noisy speech. Note that in the earlier versions, we have proposed offline estimation using noisy speech database [9] and an adaptation in each frequency bin from actual noisy speech [10]. In this study, we develop on-line estimation method in a frame-by-frame manner. The basic idea here is that we estimate and update the variance and fourth-order moment of speech spectrum and use them to match the distribution parameters. Since the estimation is carried out from noisy speech, we first describe the noise estimation. The noisy speech periodogram is smoothed as

$$\sigma_X^2\,(n,k) = \alpha_X\sigma_X^2\,(n-1,k) + (1-\alpha_X)\,|X\,(n,k)|^2\,, \tag{19}$$

where $\sigma_X^2 = \left\langle |\mathbf{X}|^2 \right\rangle$ is the variance of noisy speech spectrum, $\sigma_X^2\,(1,k) = X^2\,(1,k)$. Then the noise variance is recursively estimated using a voice activity detection (VAD)

$$\sigma_N^2\,(n,k) = \begin{cases} \alpha_N\sigma_N^2(n-1,k) + (1-\alpha_N)\sigma_X^2(n-1,k) & H0 \\ \sigma_N^2(n-1,k), & H1 \end{cases} \tag{20}$$

where $H1$ and $H0$ are the hypotheses of speech present and absent, respectively. The initial noise variance is estimated from the first 0.25 seconds duration of the observed signal

$$\widehat{\sigma_N^2\,(k)} = \frac{M}{M-1}\left(\frac{\sum_{i=1}^{M} X^2[k,i]}{M}\right), \tag{21}$$

where M is the number of initial frames. The weight factor on the right-hand side is the unbiased compensation factor for the moment estimation from a finite number of samples [15]. A VAD based on spectral distance is adopted in this study due to its possibility of it being implemented in the real-time system.

$$d\left[\vec{C}\,(n)\,,\vec{C}_N\,(n)\right] \begin{array}{ll} > \rho & H1 \\ \leq \rho & H0. \end{array} \tag{22}$$

Here $d[.]$ denotes in decibels the Euclidean distance between the current frame spectral vector $C\,(n)$ and the previously stored memory noise spectral vector; ρ is a threshold-decision. Currently, we used $\rho = 3$ dB. $H1$ and $H0$ are hypotheses described in (20). To eliminate the effect of low-frequency noises, we cut off the bins which are lower than 160 Hz before applying (22). The smoothing factors α_N $(0 < \alpha_N < 1)$ and α_X $(0 < \alpha_X < 1)$ are chosen by hearing the enhanced sound output. We experimentally found that α_N is stably good at around the interval of 0.8 – 0.98. α_X is more sensitive and should be in the interval between 0.72 – 0.78. Currently, $\alpha_N = 0.90$ and $\alpha_X = 0.75$ are used.

Given the noise estimation, we employ the decision-directed scheme [2] to estimate the signal variance and fourth-order moment. As in [2], we take a feedback from the spectral magnitude estimation in a previous frame into the estimations. For the signal variance estimation, the decision-directed estimation is denoted by

$$\begin{aligned} \sigma_S^2\,(n,k) &= \alpha_S\left|S\,(\widehat{n-1,k})\right|^2 \\ &\quad + (1-\alpha_S)\max\left[\sigma_X^2\,(n,k) - \sigma_N^2\,(n,k)\,,0\right], \end{aligned} \tag{23}$$

where α_S $(0 < \alpha_S < 1)$ is the smoothing factor, which is also chosen by experimentation. Currently, $\alpha_S = 0.9$ is used.

Analogously, for the estimation of the fourth-order moment of speech spectrum, we first estimate the fourth-order moment of noisy speech spectrum by using the conventional recursive moving average

$$\mu_{4,X}\,(n,k) = \alpha_{4,X}\mu_{4,X}\,(n-1,k) + (1-\alpha_{4,X})\,|X\,(n,k)|^4\,, \tag{24}$$

where $\mu_{4,X} = \left\langle |\mathbf{X}|^4 \right\rangle$ denotes the fourth-order moment of the noise spectrum, $\alpha_{4,X}$ $(0 < \alpha_{4,X} < 1)$ is a smoothing factor. Number 4 indicates the order of statistic. The optimal smoothing factor is also experimentally found by hearing the enhanced output sound. Currently, $\alpha_{4,X} = 0.75$ is used. The fourth-order moment of the noise spectrum is given following the Gaussian assumption, which lead to the exponential distribution in the power domain [12]

$$\mu_{4,N} = \left\langle N^4 \right\rangle = 2\sigma_N^4. \tag{25}$$

Denote the fourth-order moment of the noisy speech in terms of the noise and the speech as

$$\mu_{4,X} = \left\langle \left(S^2 + N^2 + 2SN\cos\left(\varphi_S - \varphi_N\right) \right)^2 \right\rangle. \tag{26}$$

Approximating the distribution of the phase difference by a uniform distribution, yields

$$\mu_{4,X} = \mu_{4,S} + \mu_{4,N} + 4\sigma_S^2\sigma_N^2. \tag{27}$$

The fourth-order of speech spectrum can be subtracted from noisy speech using (27). Applying one more decision-directed scheme, the estimation of the fourth-order moment of speech spectrum is expressed as

$$\begin{aligned}&\mu_{4,S}\,(n,k)\\ &\quad = \alpha_{4,S}\left|S\,(\widehat{k,n}-1)\right|^4 + (1-\alpha_{4,S})\max\\ &\qquad \times\left[\mu_{4,X}\,(n,k) - \mu_{4,N}\,(n,k) - 4\sigma_S^2\,(n,k)\,\sigma_N^2\,(n,k)\,,0\right],\end{aligned} \tag{28}$$

where $\mu_{4,S} = \left\langle |\mathbf{S}|^4 \right\rangle$ denotes the fourth-order moment of speech spectrum, $\alpha_{4,S}$ is the smoothing factor in decision-directed estimation. Currently, we use $\alpha_{4,S} = 0.9$.

The gamma model of speech power implies a relationship between $\mu_{4,S}$ and σ_S noted as

$$\mu_{4,S} = \frac{a\,(a+1)}{b^2}\sigma_S^4. \tag{29}$$

Taking into account (13) and applying one more smoothing procedure, the gamma distribution parameter estimation is given and expressed as

$$\begin{aligned}&a\,(n,k) = b\,(n,k)\\ &\quad = \alpha_a a\,(n-1,k) + (1-\alpha_a)\left(\frac{\mu_{4,S}\,(n,k)}{\sigma_S^4\,(n,k)} - 1\right)^{-1}.\end{aligned} \tag{30}$$

Note that the smoothing operator in (30) is used to remove the spikes on the estimation. We experimentally found that a small smoothing factor α_a yields a small residual noise in the enhanced signal, but gives quite a large distortion. The good compensation between noise reduction and distortion is found to be in the interval between [0.6–0.7]. Currently, $\alpha_a = 0.7$ is used. Finally, the joint distribution (2) can be determined at each time-frequency index by (15)–(18).

3. Speech Spectral Magnitude Estimation

In this section, we turn our attention to the speech spectral magnitude estimation. Given the joint distribution in (15), the MMSE and MAP estimators can be used to estimate the speech spectral magnitude. Here, we consider the generalized estimation in a general domain noted by $h(S)$. The motivation of this investigation is that the performance of system should be improved using statistical estimation in the domain, which is more close to the machine processing features or human cues. In other words, we look for a compression for the residual noise from the speech spectral magnitude estimation.

In this general domain, the MMSE estimation can be denoted as

$$h(\hat{S}) = E\left[h\,(S)\,|X\right] = \frac{\displaystyle\int_{-\infty}^{\infty} h(S)p\,(X,S)\,dS}{\displaystyle\int_{-\infty}^{\infty} p\,(X,S)\,dS}, \tag{31}$$

and the MAP estimation is expressed as

$$\hat{S} = \underset{h(S)}{\arg\max}\left[p\,(h(S)|h(X))\right]. \tag{32}$$

The MAP estimation Eq. (32) can be denoted as

$$\frac{\partial}{\partial h(S)}\left[\log\left(p\,(h(X), h(S))\right)\right] = 0, \tag{33}$$

and using the Jacobian transform, it yields the estimation equation in a compact form.

$$\frac{\partial \log\left(p\,(X,S)\right)}{\partial S} - \frac{h''\,(S)}{h'\,(S)} = 0 \tag{34}$$

We can see that, the generalized MMSE and MAP estimation Eqs. (31) and (34) return to the usual forms when $h(S) = S$. In this study, we investigate the estimations using the proposed gamma modeling in three different domains of spectral magnitude noted as SM (i.e., $h(S) = S$), generalized power domain noted as P (i.e., $h(S) = S^\alpha$) and log-spectral magnitude noted as LSM (i.e., $h(S) = \log(S)$)

3.1 Minimum Mean Square Error Estimation

3.1.1 MMSE-SM

Using (15) and (31), the MMSE estimation in the spectral magnitude domain is expressed as

$$\hat{S} = \frac{\displaystyle\int_0^{\infty} S^{2a-1}\left[S\exp\left(-\frac{S^2}{\lambda}\right)I_0\left(2S\sqrt{\frac{\vartheta}{\lambda}}\right)\right]dS}{\displaystyle\int_0^{\infty} S^{2a-2}\left[S\exp\left(-\frac{S^2}{\lambda}\right)I_0\left(2S\sqrt{\frac{\vartheta}{\lambda}}\right)\right]dS}, \tag{35}$$

where the expression inside the square brackets denotes the Rician distribution. The upper and lower terms of expression (35) then can be considered as the moments of the Rician distribution, which is evaluated in [12] and denoted in general form as

$$m\,(c,\vartheta) = \lambda^{\frac{c}{2}}\Gamma\left(\frac{c}{2}+1\right)M\left(-\frac{c}{2}+1;1;-\vartheta\right), \tag{36}$$

where $\Gamma(.)$ is the gamma function and $M(\alpha,\beta,\gamma)$ is the Kummer confluent hyper-geometrical functions [13]; ϑ is defined in (18). From (35) and (36), the MMSE estimation of speech spectral magnitude is given by

$$\hat{S} = \lambda^{\frac{1}{2}} \frac{\Gamma\left(\frac{2a+1}{2}\right)}{\Gamma(a)} \frac{M\left(-\frac{2a-1}{2};1;-\vartheta\right)}{M(-a+1;1;-\vartheta)}. \tag{37}$$

Note that, for the case $a = b = 1$, estimation (37) reproduces the well-known Ephraim-Malah estimation of speech spectral magnitude [2].

3.1.2 MMSE-P

Analogously, the MMSE estimation in the generalized power domain $h(S) = S^{\alpha}$ is given by the conditional expectation noted as

$$E[S^{\alpha}|X] = \lambda^{\frac{\alpha}{2}} \frac{\Gamma\left(\frac{\alpha}{2}+a\right) M\left(-\frac{\alpha}{2}-a+1;1;-\vartheta\right)}{\Gamma(a)\, M(-a+1;1;-\vartheta)}. \tag{38}$$

The speech spectral magnitude estimation is given as a function of a and α

$$\hat{S} = \lambda^{\frac{1}{2}} \left[\frac{\Gamma\left(\frac{\alpha}{2}+a\right) M\left(-\frac{\alpha}{2}-a+1;1;-\vartheta\right)}{\Gamma(a)\, M(-a+1;1;-\vartheta)}\right]^{\frac{1}{\alpha}}. \tag{39}$$

3.1.3 MMSE-LSM

For the MMSE estimation in the log-spectral magnitude domain, following [3], we apply the moment-generating function method, which is noted as

$$\widehat{\ln S} = \frac{\partial}{\partial\mu}\left[E\left(\exp(\mu \ln S)\,|X\right)\right]_{\mu=0}. \tag{40}$$

The moment-generating function of logarithm reproduces the moment of the joint distribution

$$\widehat{\ln S} = \frac{\partial}{\partial\mu}\left[E\left(S^{\mu}|X\right)\right]_{\mu=0}. \tag{41}$$

The right term of (41) is given in (38) and differentiating it by terms yields

$$\begin{aligned}\frac{\partial}{\partial\mu}E\left[S^{\mu}|X\right]_{\mu=0} &= \ln\lambda + \frac{\frac{\partial}{\partial\mu}\Gamma\left(\frac{\mu}{2}+a\right)_{\mu=0}}{2\Gamma(a)} \\ &\quad + \frac{\frac{\partial}{\partial\mu}M\left(-\frac{\mu}{2}-a+1;1;-\vartheta\right)_{\mu=0}}{2M(-a+1;1;-\vartheta)}.\end{aligned} \tag{42}$$

The two components on the right-hand of (42) can be denoted as

$$\frac{\frac{\partial}{\partial\mu}\Gamma\left(\frac{\mu}{2}+a\right)_{\mu=0}}{\Gamma(a)} = \psi(a), \tag{43}$$

$$\begin{aligned}&\frac{\frac{\partial}{\partial\mu}M\left(-\frac{\mu}{2}-a+1;1;-\vartheta\right)_{\mu=0}}{M(-a+1;1;-\vartheta)} \\ &= \frac{\sum_{k=0}^{\infty} \frac{(-a+2)_{k-1}}{k!}\frac{z^k}{k!}}{M(-a+1;1;-\vartheta)} - \psi(-a+2),\end{aligned} \tag{44}$$

where $\psi(.)$ is a polygamma function [13]. For the case $a = b = 1$, (42) can be simplified by an approximation, it reproduces the well-known Ephraim-Malah's estimation for the conventional Gaussian model [3]

$$\begin{aligned}\widehat{\ln S} &= \ln\lambda + \psi(1) + \frac{1}{2}\sum_{k=0}^{\infty}\frac{1}{k}\frac{(-\vartheta)^k}{k!} \\ &= \ln\lambda + \frac{1}{2}\left(\ln\vartheta + \int_{\vartheta}^{\infty}\frac{e^t}{t}dt\right)\end{aligned} \tag{45}$$

We can see that, the MMSE estimations tend to more complicated using general prior distribution. Naturally, an appropriate approximations of (37), (39) and (42) might reduce the computational cost and get these methods to be realized in the real systems. Unfortunately, this is not simple mathematical problem, and we remain it to the future work. In this study, we implement the estimations (37), (39) and (42) using the numerical calculations of the hypergeometrical function [13].

3.2 Maximum a Posterior Probability Estimation

Maximum a posterior probability estimation is a powerful estimation method, which requires the prior information. Since the main point of this study is the improvement of the prior distribution modeling of speech spectrum, this estimation method is suitable from the theoretical point of view. We will show that the MAP estimation is also suitable for the implementation.

3.2.1 MAP-SM

Denote the Eq. (34) for the case of the estimation in spectral magnitude domain as

$$\frac{\partial}{\partial S}\left[\log\left(p(X|S)\right)\right] + \frac{\partial}{\partial S}\left[\log\left(p(S)\right)\right] = 0. \tag{46}$$

Here, the Bessel function is approximated by [13]

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}}e^{x}, \quad x > 0. \tag{47}$$

Note that, the relative error given by the approximation (47) is less than 1% at $x > 0.2$, i.e.,

$$\frac{2XS}{\sigma_N^2} > 0.2. \tag{48}$$

When speech magnitude is small,

$$\frac{2X}{\sigma_N^2} \approx \frac{2S}{\sigma_N^2} \approx \frac{2\sigma_S}{\sigma_N^2}, \tag{49}$$

substituting (48) into (49) yields a constraint for the local SNR denoted as $10\log_{10}\frac{\sigma_S^2}{\sigma_N^2} > -20\,\text{dB}$, and therefore, the approximation error is negligible.

Using (47), the first component in (46) can be expressed as

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right)\right] = -\frac{2S}{\sigma_N^2} - \frac{1}{2S} + \frac{2X}{\sigma_N^2}. \tag{50}$$

For the proposed gamma model of speech power, substituting (14) and (50) into (46), yields a second-order equation for the gain function $G = \frac{S}{X}$ as

$$-G^2 + \frac{G}{\left(1 + \frac{b}{\xi}\right)} + \frac{4a - 3}{4\gamma\left(1 + \frac{b}{\xi}\right)} = 0, \tag{51}$$

where ξ and γ are defined in (18). A closed-form solution is obtained and denoted by

$$G = \frac{1}{2\left(1 + \frac{b}{\xi}\right)} + \sqrt{\frac{1}{4\left(1 + \frac{b}{\xi}\right)^2} + \frac{4a - 3}{4\gamma\left(1 + \frac{b}{\xi}\right)}}. \tag{52}$$

The closed-form solution for the MAP estimation is important because we can exactly yield the global maximum of the posterior probability. Moreover, this tractable solution is suitable for the implementation. Note that, the MAP solution (52) is generalized of the solution given by using the conventional Gaussian model derived by Wolfe and Godsill [4].

3.2.2 MAP-P

Using (34), the MAP estimation equation in the generalized power domain $h(S) = S^{\alpha}$ is given as

$$\frac{\partial \log\left(p\left(X, S\right)\right)}{\partial S} - \frac{(\alpha - 1)}{S} = 0. \tag{53}$$

Analogously, the gain function is given in a closed-form solution as

$$G = \frac{1}{2\left(1 + \frac{b}{\xi}\right)} + \sqrt{\frac{1}{4\left(1 + \frac{b}{\xi}\right)^2} + \frac{4a - 2\alpha - 1}{4\gamma\left(1 + \frac{b}{\xi}\right)}}. \tag{54}$$

One interesting result is that, the Wiener filter can be considered as a special case of (54), when $a = b = 1$ (i.e., the conventional Gaussian model) and $\alpha = 1.5$. The gain function in that case is given as

$$G = \frac{\sigma_S^2}{\sigma_N^2 + \sigma_S^2} = \left(1 + \frac{1}{\xi}\right)^{-1}. \tag{55}$$

3.2.3 MAP-LSM

The MAP estimation equation for the estimation in the log-spectral magnitude domain is derived and expressed as

$$\frac{\partial \log\left(p\left(X, S\right)\right)}{\partial S} + \frac{1}{S} = 0. \tag{56}$$

Analogously, the gain function is given in a closed-form extraction as

$$G = \frac{1}{2\left(1 + \frac{b}{\xi}\right)} + \sqrt{\frac{1}{4\left(1 + \frac{b}{\xi}\right)^2} + \frac{4a - 1}{4\gamma\left(1 + \frac{b}{\xi}\right)}}. \tag{57}$$

The MAP-LSM gain function for the conventional Gaussian model is given as a special case of (57) noted by

$$G = \frac{1}{2\left(1 + \frac{1}{\xi}\right)} + \sqrt{\frac{1}{4\left(1 + \frac{1}{\xi}\right)^2} + \frac{3}{4\gamma\left(1 + \frac{1}{\xi}\right)}}. \tag{58}$$

3.3 Gain Curves

From the results given in previous paragraph, it can be seen that the MMSE and MAP gain function based on the proposed gamma modeling are controlled by the prior distribution parameters and the conventional Gaussian model is a special case when these parameters are fixed as $a = b = 1$. Since the performance of the enhancement filter is related to the curve of the gain function, in this paragraph, we do the investigation of these curves. Figures 4 and 5 show the gain functions based on gamma modeling with two different prior distribution parameters. Thegain functions based on Gaussian model are also plotted as references. These gain functions are functions of the instantaneous SNR, defined by $(\gamma - 1)$ [2]. From these figures, we can see that for the Gaussian model, the Wiener filter and the MAP-LSM estimation produce the highest and lowest reduction levels, respectively. This effect explains the fact that the Wiener filter often over reduces the noise at the speech beginning and ending frames. The curves of the gain functions of the gamma model are various by the distribution parameter. When the gamma parameter is small, the gamma model yields a lower reduction level than the Wiener filter under a low-input SNR

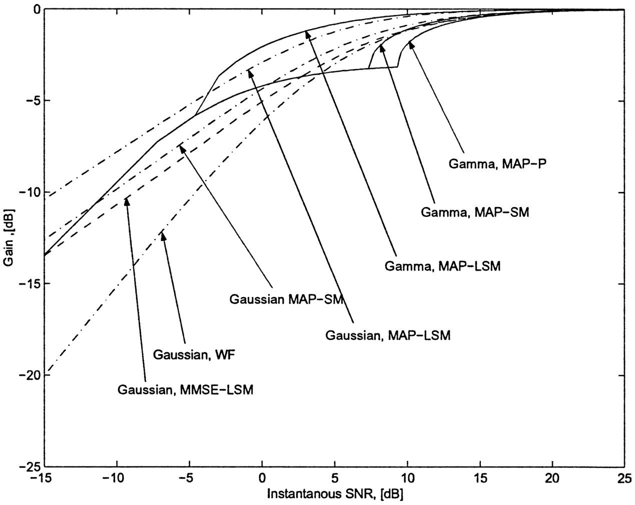

Fig. 4 Gain curve for $a = b = 0.1$.

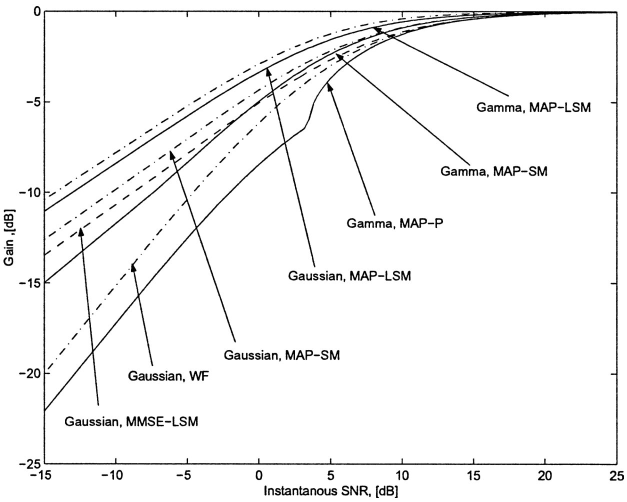

Fig. 5 Gain curve for $a = b = 0.8$.

and a higher reduction level under a high-input SNR. During vowel duration and in near-formant frequency bins, the speech spectral distribution is less sparse and therefore, the gamma distribution parameter must be larger. In this case, the gamma model maintains the lower reduction level and provides less distortions. Since the gamma distribution parameters are adapted from actual noisy speech, this effect can be considered as an automatic optimization of the gain function by the prior modeling and in the next section we confirm this consideration by the experiments.

4. Experiments

4.1 Speech Enhancement Implementation

The speech spectral magnitude estimation methods using the proposed gamma modeling are implemented in speech enhancement systems. A diagram of processing is shown in Fig. 6. The noisy speech is transformed into the STDFT domain using a hamming window with a frame length of 25 ms and a frame shift of 10 ms. The noise and signal statistics (i.e., the variance and fourth-order moment) are estimated in order to adapt the gamma parameters. The adaptive gamma distribution parameters are used in the MMSE and MAP speech spectral magnitude estimations according to (37), (39), (42), (52), (54) and (57). The systems are named by the estimation method and the domain to be applied. They are MMSE-SM, MMSE-P, MMSE-LSM, MAP-SM, MAP-P and MAP-LSM, respectively. Note that, we do not consider the optimization of the generalized power order and it was chosen by the experimentation, currently $\alpha = 1.5$ is used. For the reference, the similar estimations are implemented using the conventional Gaussian model (i.e., $a = b = 1$). The Gaussian-model-based systems are identified by adding the symbol G. Finally, the phase adding, and "overlap and add" techniques are applied to synthesis the enhanced signal.

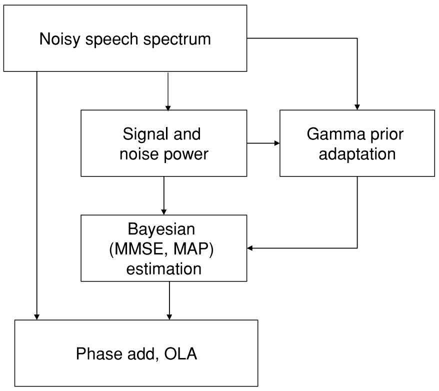

Fig. 6 Statistical speech enhancement using the proposed gamma model.

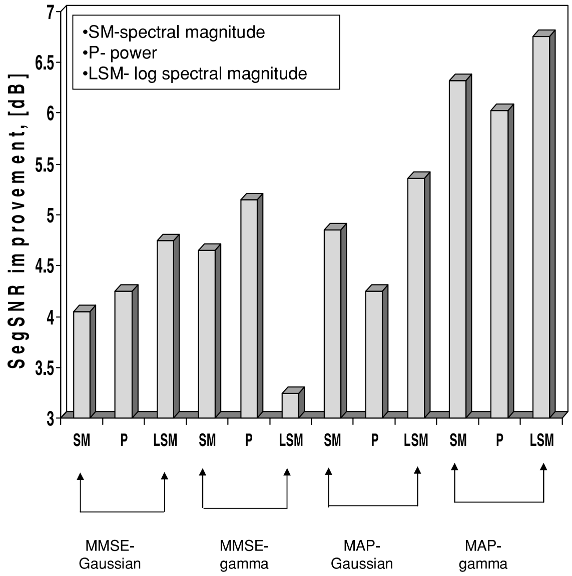

Fig. 7 Overall results of segmental SNR improvement [dB].

4.2 Evaluation and Discussion

The proposed speech enhancement systems are tested using the AURORA2J database [17]. The overall results of the segmental SNR improvements is shown in Fig. 7 and the relative improvements of automatic speech recognition (ASR) for the clean training and multi-conditions training are in Figs. 8 and 9. For automatic speech recognition experiments, we apply enhancement filters to both testing and training databases. The digit HMMs are standard complex back-end models of 16 states, and each state has a 20-component GMM with a diagonal covariance matrix [16]. From these figures, it can be seen that the proposed gamma model performs better than the conventional Gaussian model. This can be explained by the fact that the performances of MMSE and MAP estimations are dependent on the accuracy in the speech prior distribution modeling. Using more general distribution with frame-by-frame parameter estimation, the proposed method provides more accurate modeling and this is the principal merit of the method and

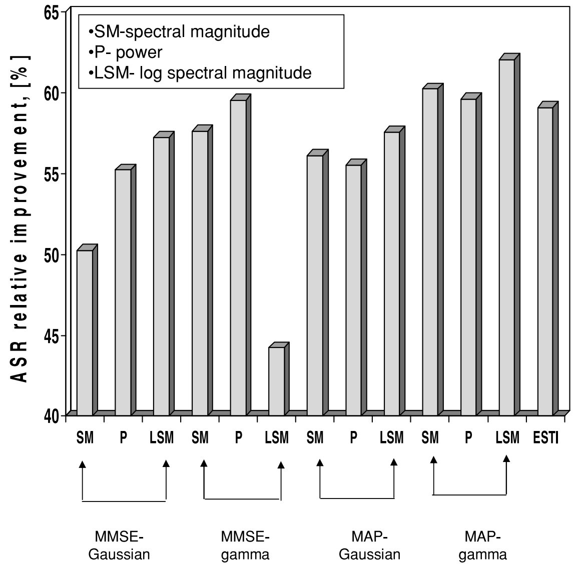

Fig. 8 Overall results of ASR evaluation using AURORA-2J database in relative improvement for clean training [%].

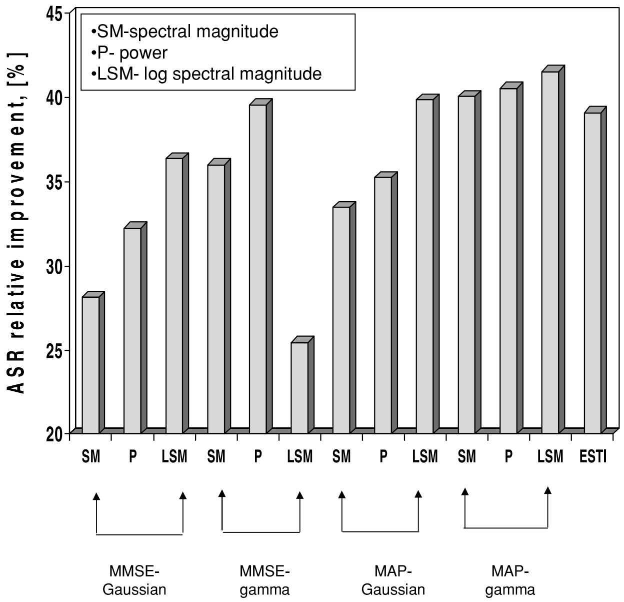

Fig. 9 Overall results of ASR evaluation using AURORA-2J database in relative improvement for multi-conditions training [%].

Table 1 Listening test: Q1-Which one is less distorted? Q-2 Which one is less noisy? Q-3 Which one is best?

Q	Subway	Babble	Car	Exhibition
1	MAP-LSM	MAP-LSM	MMSE-P	MAP-LSM
2	MAP-P	MAP-P	MAP-P	MAP-P
3	MAP-LSM	MAP-SM	MAP-LSM	MAP-SM

Q	Restaurant	Street	Airport	Station
1	MAP-LSM(G)	MAP-LSM	MAP-SM	MAP-LSM
2	MAP-P	MAP-SM	MAP-P	MAP-P
3	MMSE-LSM(G)	MMSE-LSM	MAP-SM	MAP-LSM

Table 2 Best ASR evaluation results using AURORA2J database under each noise condition.

Noise	Subway	Babble	Car	Exhibition
Method	MAP-LSM	MAP-LSM	MAP-LSM	MAP-SM

Noise	Restaurant	Street	Airport	Station
Method	MMSE-P	MAP-SM	MAP-LSM	MAP-LSM

the reason for the better performance. Among the methods based on the proposed gamma model, the MAP estimations yield better results than the MMSE estimations. For the MMSE estimations, though the MMSE-SM and MMSE-P give better results than the similar ones based on the conventional Gaussian model, the computational cost is much more expensive. The performance inferior of MMSE-LSM can be explained by the fact that the errors occur in the numerical calculations of the derivatives of the hyper-geometrical functions. Note that, although this performance degradation is caused by the implementation problem but not theoretical problem, the MMSE estimations tend to be more complex than the MAP estimation using a more general prior distribution and therefore is not recommended. The MAP estimations based on the proposed gamma modeling are given in closed-form solutions and from our experience the computational cost of these methods are approximately the same as the Ephraim-Malah's MMSE-LSM method. The MAP-LSM performance is the best with approximately 2 dB of the segmental SNR improvement, 5% of the relative improvement in the clean training and 4% in the multi-conditions training compared to the Ephraim-Malah's MMSE-LSM. The systems based on MAP estimations using the proposed gamma modeling overcome the conventional ESTI front-end [18], which is specially designed for speech recognition but does not provide enhanced signals. The proposed method provide improvements in both ASR performance and the sound quality and therefore is more appropriated for the communication applications. A simple listening test is performed by four subjects listening to 25 randomly chosen utterances of each noise type. Table 1 shows the results of the listening test. One interesting fact is that the MAP-P estimation using the gamma model is the best method for the noise reduction as the MAP-LSM estimation provides less distortion and is considered to have the best performance under most noise conditions. Finally, Table 2 shows the best ASR evaluation results under each noise environment. We can see that, the MAP-LSM estimation using the proposed gamma model is the best under 6 from 8 noise environments. The MMSE estimation perform a little better than the MAP estimation only under the restaurant noise environment. In both cases, the gamma-model-based systems perform better than those based on the conventional Gaussian model. The superiority of the MAP compared to the MMSE using the proposed gamma model can be explained by follows. The MAP estimation is more sensitive to the prior distribution than the MMSE estimation. The proposed gamma distribution of speech power and its on-line estimation particularly improved the prior modeling and therefore this improvement was better realized in the MAP filters. The best performance of the estimation in the log-spectral magnitude domain confirm the fact that this domain is close to the human cues and speech recognition feature.

5. Conclusions

The main points of this study are summarized as follows. First, we propose a more general and flexible distribution for the speech spectrum distribution modeling. Second, we propose an adaptive estimation of the modeled distribution parameters from actual noisy speech. Third, we derive, implement and investigate the MMSE and MAP filters in different domains using the proposed model. The MMSE estimations tend to complicated using more general prior distributions and therefore are not recommended. In contrast, the MAP estimations are suitable for the implementation and yield better performance in both speech recognition and sound quality. The MAP estimation in log-spectral magnitude and generalized power domains are recommended. The optimal choice of the power order α is an interesting problem what we remain in the future work. We also intend to do more experimental evaluations in other objective measurements on the sound quality of the enhanced signals. This model might also be applied for other signals such as music or biomedical sounds.

References

[1] R. Martin, "Statistical methods for enhancement of noisy speech," Proc. IWAENC, Kyoto, 2003.

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

[3] Y. Ephraim and D. Malah, "Speech enhancement using MMSE log-spectral amplitude estimations," IEEE Trans. Acoust. Speech Signal Process., vol.33, no.2, pp.443–445, 1985.

[4] P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim suppression rule for speech enhancement," IEEE Workshop on Statistical Signal Processing, 2001.

[5] T. Lotter and P. Vary, "Noise reduction by maximum a posteriori spectral amplitude estimation with super-Gaussian speech modeling," Proc. IWAENC, Kyoto, 2003.

[6] R. Martin, "Speech enhancement using MMSE short-time spectral estimation with gamma speech prior," Proc. ICASSP 02, Orlando, FL, 2002.

[7] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, 1993.

[8] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," IEEE Trans. Acoust. Speech Signal Process., vol.10, no.6, pp.341–351, 2002.

[9] T.H. Dat, K. Takeda, and F. Itakura, "Speech enhancement based on magnitude estimation using the gamma prior," Proc. ICSPL, Jeju, Korea, 2004.

[10] T.H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," Proc. ICASSP, Philadelphia, PA, 2005.

[11] T.H. Dat, K.Takeda, and F. Itakura, "The MAP and cumulative distribution equalization methods for speech spectral estimation," Proc. ISCA ITRW NOLISP05, Barcelona, Spain, 2005.

[12] E. Pazen, Modern Probability Theory and Its Applications, Wiley, 1992.

[13] S. Zhang and J. Jin, Computation of Special Functions, Wiley, 1996.

[14] R. Martin, "Noise power spectral estimation based on optimal smoothing and minimum statistics," IEEE Trans. Acoust. Speech Signal Process., vol.9, no.5, pp.504–512, 2001.

[15] C. Rose and M.D. Smith, "k-Statistics: Unbiased estimators of cumulants," in Mathematical Statistics with Mathematica, pp.256–259, Springer-Verlag, New York, 2002.

[16] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. ISCA ITRW ASR, 2000.

[17] http://sp.shinshu-u.ac.jp/CENSREC

[18] ESTI standard document, ETSI ES201 108 v1.1.2 (2000-04), 2000.

Tran Huy Dat was born in Hanoi, Vietnam, in 1971. He received the Master of Engineering degree in 1995 from the Ukrainian National Technical University in 1995. In 2000, he received the PhD of Physic-Mathematical Science degree from the National Academy of Sciences of Ukraine. From 2000 to 2002 he did his postdoc research at the Institute of Hydromechanics, National Academy of Science of Ukraine. From 2002 to 2005 he was a postdoc researcher at the Itakura and Takeda Labs, Nagoya University. He presently works as a Scientist at the Institute for Infocomm Research, Singapore. His research interest is including acoustic scattering, wave field modeling, statistical signal processing, speech enhancement, and speech recognition. He is a member of the Acoustical Society of America.

Kazuya Takeda received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1983, 1985, and 1994 respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in a project for constructing voice-activated telephone extension system. He has joined Graduate School of Nagoya University in 1995. Since 2003, he is a professor at Graduate School of Infomation Science at Nagoya University. He is a member of the IEEE and the ASJ.

Fumitada Itakura was born in Toyokawa near to Nagoya, in 1940. He earned undergraduate and graduate degrees at Nagoya University. In 1968, he joined NTT's Electrical Communication Laboratory in Musashino, Tokyo. He completed his Ph.D. in speech processing in 1972, writing his dissertation on "Speech Analysis and Synthesis System based on a Statistical Method." He worked on isolated word recognition in the Acoustics Research Department of Bell Labs under James Flanagan from 1973 to 1975. Between 1975 and 1981, he researched problems in speech analysis and synthesis based on the Line Spectrum Pair [LSP] method. In 1981, he was appointed as Chief of the Speech and Acoustics Research Section at NTT. He left this position in 1984 to take a professorship in communications theory and signal processing at Nagoya University. After 20 years of teaching and research at Nagoya University, he retired from Nagoya University and joined Meijo University in Nagoya. His major contributions include theoretical advances involving the application of stationary stochastic process, linear prediction, and maximum likelihood classification to speech recognition. He patented the PARCOR vocoder in 1969 the LSP in 1977. His awards include the IEEE ASSP Senior Award, 1975, an award from Japan's Ministry of Science and Technology, 1977, the 1986 Morris N. Liebmann Award (with B.S. Atal), the 1997 IEEE Signal Processing Society Award, and the IEEE third millennium medal. He is a fellow of the IEEE, and a member of the Acoustical Society Japan.