

LETTER

Speech Prior Estimation for Generalized Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator

Ryo WAKISAKA^{†a)}, *Nonmember*, Hiroshi SARUWATARI[†], *Member*, Kiyohiro SHIKANO[†], *Fellow*,
and Tomoya TAKATANI^{††}, *Member*

SUMMARY In this paper, we introduce a generalized minimum mean-square error short-time spectral amplitude estimator with a new prior estimation of the speech probability density function based on moment-cumulant transformation. From the objective and subjective evaluation experiments, we show the improved noise reduction performance of the proposed method.

key words: *generalized MMSE STSA estimator, speech kurtosis estimation, cumulant*

1. Introduction

In recent years, many applications of speech communication systems have been developed and launched into real-world human interface. In such applications, the essential requirement of these systems is robustness against environmental noise to work stably under adverse noise conditions. Therefore, many nonlinear noise reduction methods, such as spectral subtraction, Wiener filtering, and the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator, have been actively studied [1].

The basic theory of the MMSE STSA estimator has been presented by Ephraim et al. [2] for the optimal identification of target speech amplitude spectrum in the MMSE sense, ignoring phase spectrum information. In this paper, we mainly deal with the MMSE STSA estimator because of its good balance of noise reduction ability and less distortion property in speech. The original MMSE STSA estimator has, however, an inherent problem owing to the mismatch in speech prior assumption. Thus, the method assumes that the speech probability density function (p.d.f.) has a Gaussian distribution. In fact, however, the speech signal always obeys more spiky p.d.f. [3] and this mismatch is likely to cause a degradation in enhanced speech quality.

To cope with this mismatch problem, an improved MMSE STSA estimator, which assumes that the speech amplitude spectrum obeys the generalized gamma distribution [4], *generalized MMSE STSA estimator* [5], [6], has been proposed. However, in Refs. [5], [6], no method of estimating the shape parameter of the speech p.d.f. in advance is discussed, and this still remains as an open problem. The

difficulty in the speech prior estimation problem is that the speech component is always overlapped with noise at every time-frequency grid; the speech p.d.f. is identified by the higher-order moments of speech amplitude spectrum but these moments do not hold the additivity for additive variables.

In this paper, to enable the generalized MMSE STSA estimator to treat various types of speech signals with different p.d.f., we propose a new algorithm for the prior estimation of the hidden speech p.d.f. even under the noise-contaminated condition. Our method is based on a moment-cumulant transformation technique with respect to the statistical estimates of observable noise and noisy speech signals. The cumulant-domain calculation resolves the signal-additivity problem, and also the moment-domain calculation deals well with the statistics transformation from the waveform to the amplitude spectrum. From the objective and subjective evaluation experiments, we show the improved noise reduction performance of the proposed method.

2. Conventional Method

2.1 Original MMSE STSA Estimator and Its Problem

We consider an acoustic mixing model where the observed signal contains only one target speech signal, and an additive noise signal. Hereafter, the observed signal in the time-frequency domain, $x(f, \tau)$, is given by

$$x(f, \tau) = s(f, \tau) + n(f, \tau), \quad (1)$$

where f is the frequency bin number, τ is the time-frame index number, $s(f, \tau)$ is target speech signal component, and $n(f, \tau)$ is the additive noise signal. Ephraim et al. [2] have proposed the noise reduction method that estimates the speech amplitude spectrum on the basis of MMSE criterion under a certain (fixed) speech prior; hereafter, we call this Ephraim method *original MMSE STSA estimator*. The original (nonparametric) MMSE STSA estimator assumes that the speech signal obeys a Gaussian distribution. However, it is well known that the speech signal has more spiky p.d.f. similar to a Laplacian distribution. Therefore, we introduce the generalized (parametric) MMSE STSA estimator with a new prior estimation of the speech p.d.f. Figure 1 shows the block diagram of our proposed method.

Manuscript received August 10, 2011.

Manuscript revised October 17, 2011.

[†]The authors are with Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan.

^{††}The author is with Toyota Motor Corporation, Toyota-shi, 470-0309 Japan.

a) E-mail: ryo-w@is.naist.jp

DOI: 10.1587/transfun.E95.A.591

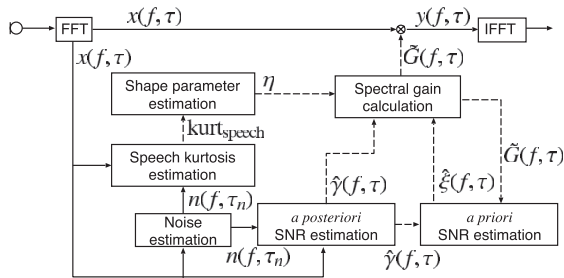


Fig. 1 Block diagram of proposed method.

2.2 Parametric Model for Speech

In this study, we introduce the *generalized gamma distribution* [4] to model the amplitude spectral grid signal in the time-frequency domain. Its p.d.f. is written as

$$p(x) = 2\phi^\eta \Gamma(\eta)^{-1} x^{2\eta-1} \exp(-\phi x^2), \quad (2)$$

$$\phi = \eta / E\{|x|^2\}, \quad (3)$$

where $\Gamma(\cdot)$ denotes the gamma function and η ($0 < \eta < 1$) is a shape parameter; $\eta = 1$ gives a Rayleigh distribution that corresponds to a Gaussian signal and a smaller value of η corresponds to a super-Gaussian signal. Note that strict proof on justification of this statistical speech model is out of this paper's scope, but we add an experimental evaluation on this issue in Appendix.

2.3 Generalized MMSE STSA Estimator [5]

The generalized MMSE STSA estimator can estimate temporal a priori and a posteriori SNRs and the spectral gain using the noise signal observed in the nonspeech period. First, the a posteriori SNR estimate $\hat{\gamma}(f, \tau)$ is given as

$$\hat{\gamma}(f, \tau) = |x(f, \tau)|^2 / \lambda(f) = |x(f, \tau)|^2 / E\{|n(f, \tau)|^2\}, \quad (4)$$

where $\lambda(f)$ is the power spectrum of the observed noise, and $E\{\cdot\}$ denotes the expectation operator.

Next, using Eq. (4), the a priori SNR estimate $\hat{\xi}(f, \tau)$ is given as [2]

$$\hat{\xi}(f, \tau) = \alpha \hat{\gamma}(f, \tau - 1) \tilde{G}^2(f, \tau - 1) + (1 - \alpha) P[\hat{\gamma}(f, \tau - 1)], \quad (5)$$

where α is the weighting factor of the decision-directed estimation, $\tilde{G}(f, \tau)$ is a spectral gain function, and the operator $P[\cdot]$ is a flooring function in which the negative input is floored to zero. Also, the spectral gain function is defined as [5]

$$\tilde{G}(f, \tau) = \frac{\sqrt{\tilde{\nu}(f, \tau)} \Gamma(\eta + 0.5) \Phi(0.5 - \eta, 1, -\tilde{\nu}(f, \tau))}{\hat{\gamma}(f, \tau) \Gamma(\eta) \Phi(1 - \eta, 1, -\tilde{\nu}(f, \tau))}, \quad (6)$$

$$\tilde{\nu} = \hat{\xi}(f, \tau) (\eta + \hat{\xi}(f, \tau))^{-1} \hat{\gamma}(f, \tau),$$

where Φ is a confluent hypergeometric function. The gain $\tilde{G}(f, \tau)$ includes a shape parameter η that should represent

speech p.d.f. prior, and we discuss how to determine it optimally in Sect. 3.

Finally, noise reduction is carried out as follows:

$$y(f, \tau) = \tilde{G}(f, \tau) x(f, \tau), \quad (7)$$

where $y(f, \tau)$ is the resultant output signal.

3. Estimation of Optimal Shape Parameter of Speech p.d.f.

3.1 Shape Parameter and Kurtosis

Regarding the gamma distribution $p(x)$ in Eq. (2), it is well known that the shape parameter η can be written as

$$\eta = (\mu_4 / \mu_2^2 - 1)^{-1}, \quad (8)$$

where μ_4 / μ_2^2 is called the *kurtosis* and μ_m is the m th-order moment of the amplitude spectrum. From this relation, the shape parameter of the subjective speech signal can be estimated by obtaining its kurtosis value. In general, however, it is difficult to directly estimate the kurtosis of a speech signal because of the contamination by additive noise. In this section, a new algorithm of speech kurtosis estimation is proposed for the estimation of the shape parameter of speech p.d.f.

3.2 Problem and Strategy

Since the speech component is always contaminated with noise at every time-frequency grid, it is difficult to estimate the speech kurtosis via theoretical analysis. Therefore, we inversely calculate the kurtosis of the speech amplitude spectrum in a data-driven manner, utilizing two observable statistics of the noisy speech signal and noise signal estimated in the speech-absent part. Note that the proposed speech kurtosis estimation is still an unsupervised method because this method requires no reference (clean) speech signals.

To cope with the mathematical problem that the mixing of speech and noise is additive but generally their higher-order moments are not additive, we introduce the *cumulant*, which holds the additivity for additive variables. Meanwhile, in transformation from a waveform to its amplitude spectrum, the exponentiation operation is conducted but the cumulant does not have a straightforward relationship. In this case, we use the moment instead of the cumulant. Thus, we propose to use *moment-cumulant transformation*.

3.3 Moment-Cumulant Transformation

In this section, we derive some formula regarding moment-cumulant transformation. They explicitly represent the relations between the moment and cumulant in each order, which are useful for estimating the kurtosis of the speech amplitude spectrum.

First, the characteristic function $\phi_x(it)$ of the random

variable x is defined as

$$\phi_x(it) = \int_{-\infty}^{\infty} e^{itx} P(x) dx. \quad (9)$$

Then, we can define the m th-order moment $\mu_m(x)$ and the m th-order cumulant $\kappa_m(x)$ of x as follows:

$$\mu_m(x) = \left. \frac{\partial^{(m)} \phi_x(it)}{\partial it^{(m)}} \right|_{t=0}, \quad (10)$$

$$\kappa_m(x) = \left. \frac{\partial^{(m)} \log \phi_x(it)}{\partial it^{(m)}} \right|_{t=0}. \quad (11)$$

Next, polynomial forms of interrelations between the moment and cumulant are derived below. From Eq. (10), the m th-order moment $\mu_m(x)$ can be rewritten as

$$\begin{aligned} \mu_m(x) &= \left. \frac{\partial^{(m)} \exp(\log \phi_x(it))}{\partial it^{(m)}} \right|_{t=0} \\ &= \sum_{\pi(m)} \exp^{(|\pi(m)|)} (\log \phi_x(it)) \\ &\quad \prod_{B \in \pi(m)} [\log \phi_x(it)]^{(|B|)} \Big|_{t=0} \\ &= \sum_{\pi(m)} \prod_{B \in \pi(m)} \kappa_{|B|}(x), \end{aligned} \quad (12)$$

where we use a *combinational form of Faà di Bruno's formula*,

$$\frac{\partial^{(m)} f(g(x))}{\partial x^{(m)}} = \sum_{\pi(m)} f^{(|\pi(m)|)}(g(x)) \prod_{B \in \pi(m)} [g(x)]^{(|B|)}, \quad (13)$$

where $\pi(m)$ runs through the list of all partitions of a set of size m , $B \in \pi(m)$ means that B is one of the blocks into which the set is partitioned, and $|B|$ is the size of the set B .

In the same manner, from Eq. (11), the m th-order cumulant $\kappa_m(x)$ is given by

$$\begin{aligned} \kappa_m(x) &= \sum_{\pi(m)} \log^{(|\pi(m)|)}(\phi_x(it)) \prod_{B \in \pi(m)} [\phi_x(it)]^{(|B|)} \Big|_{t=0} \\ &= \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(x). \end{aligned} \quad (14)$$

3.4 Estimation of Speech Kurtosis from Observations

Hereafter, we define complex-valued variables of the observed (noisy speech) signal, the original speech signal, and the noise signal as $(x_R + ix_I)$, $(s_R + is_I)$, and $(n_R + in_I)$, respectively, where $x_R = s_R + n_R$ and $x_I = s_I + n_I$ hold. Only the statistics of $(x_R + ix_I)$ and $(n_R + in_I)$ are observable, but that of $(s_R + is_I)$ is a hidden value to be estimated. First, we measure the following m th-order moments from data;

$$\mu_m(x_R) = E[x_R^m], \quad (15)$$

$$\mu_m(x_I) = E[x_I^m], \quad (16)$$

$$\mu_m(n_R) = E[n_R^m], \quad (17)$$

$$\mu_m(n_I) = E[n_I^m], \quad (18)$$

where x_R and x_I are obtained directly from whole samples of the observed signal. Also, n_R and n_I can be measured in a speech-pause time period, which is assumed to be obtained by an appropriate method such as voice activity detector in our study.

Generally, the cumulant has the additivity for the additive independent variables, i.e., $\kappa_m(a + b) = \kappa_m(a) + \kappa_m(b)$. Using this relation and Eq. (14), we can estimate the cumulant of the real part of the speech signal as

$$\begin{aligned} \kappa_m(s_R) &= \kappa_m(x_R) - \kappa_m(n_R) \\ &= \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(x_R) \\ &\quad - \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(n_R). \end{aligned} \quad (19)$$

In the same manner, the cumulant of the imaginary part is written as

$$\begin{aligned} \kappa_m(s_I) &= \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(x_I) \\ &\quad - \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(n_I). \end{aligned} \quad (20)$$

Next, the statistics of the squared variable of s_R is given by

$$\mu_m(s_R^2) = \mu_{2m}(s_R) = \sum_{\pi(2m)} \prod_{B \in \pi(2m)} \kappa_{|B|}(s_R). \quad (21)$$

In the imaginary part, $\mu_m(s_I^2)$ is also written as

$$\mu_m(s_I^2) = \mu_{2m}(s_I) = \sum_{\pi(2m)} \prod_{B \in \pi(2m)} \kappa_{|B|}(s_I). \quad (22)$$

Given Eqs. (21) and (22), we can calculate the cumulant of the power spectrum $s_R^2 + s_I^2$ as

$$\begin{aligned} \kappa_m(s_R^2 + s_I^2) &= \kappa_m(s_R^2) + \kappa_m(s_I^2) \\ &= \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(s_R^2) \\ &\quad + \sum_{\pi(m)} (-1)^{|\pi(m)|-1} (|\pi(m)| - 1)! \prod_{B \in \pi(m)} \mu_{|B|}(s_I^2), \end{aligned} \quad (23)$$

and the m th-order moment of the power spectrum is given by

$$\mu_m(s_R^2 + s_I^2) = \sum_{\pi(m)} \prod_{B \in \pi(m)} \kappa_{|B|}(s_R^2 + s_I^2). \quad (24)$$

Furthermore, the m th-order moment of the amplitude spectrum $(s_R^2 + s_I^2)^{\frac{1}{2}}$ is

$$\mu_m((s_R^2 + s_I^2)^{\frac{1}{2}}) = \mu_m(s_R^2 + s_I^2). \quad (25)$$

Finally, using Eqs. (15)–(25), we can estimate the resultant kurtosis of the speech amplitude spectrum as

$$\begin{aligned} \text{kurt}_{\text{speech}} &= \frac{\mu_4((s_R^2 + s_I^2)^{\frac{1}{2}})}{\mu_2^2((s_R^2 + s_I^2)^{\frac{1}{2}})} \\ &= \frac{\mathcal{N}(\mu_m(x_R), \mu_m(x_I), \mu_m(n_R), \mu_m(n_I))}{\mathcal{D}(\mu_m(x_R), \mu_m(x_I), \mu_m(n_R), \mu_m(n_I))}, \end{aligned} \quad (26)$$

where

$$\begin{aligned} &\mathcal{N}(\mu_m(x_R), \mu_m(x_I), \mu_m(n_R), \mu_m(n_I)) \\ &= \mu_4(x_R) + \mu_4(x_I) - \mu_4(n_R) - \mu_4(n_I) \\ &\quad + 6\mu_2^2(n_R) + 6\mu_2^2(n_I) + 2\mu_2(x_R)\mu_2(x_I) + 2\mu_2(n_R)\mu_2(n_I) \\ &\quad - 6\mu_2(x_R)\mu_2(n_R) - 6\mu_2(x_I)\mu_2(n_I) \\ &\quad - 2\mu_2(x_R)\mu_2(n_I) - 2\mu_2(x_I)\mu_2(n_R), \end{aligned} \quad (27)$$

$$\begin{aligned} &\mathcal{D}(\mu_m(x_R), \mu_m(x_I), \mu_m(n_R), \mu_m(n_I)) \\ &= \mu_2^2(x_R) + \mu_2^2(x_I) + \mu_2^2(n_R) + \mu_2^2(n_I) + 2\mu_2(x_R)\mu_2(x_I) \\ &\quad - 2\mu_2(x_R)\mu_2(n_R) - 2\mu_2(x_R)\mu_2(n_I) - 2\mu_2(x_I)\mu_2(n_R) \\ &\quad - 2\mu_2(x_I)\mu_2(n_I) + 2\mu_2(n_R)\mu_2(n_I). \end{aligned} \quad (28)$$

The shape parameter of speech p.d.f. can be estimated using Eqs. (8) and (26).

4. Experiment

4.1 Experimental Setup

We conducted experiments to confirm the effectiveness of the proposed method. In this experiment, we compare the original MMSE STSA estimator and the generalized MMSE STSA estimator with the proposed speech p.d.f. estimation (hereafter, we simply refer to this method as *generalized MMSE STSA estimator*).

We used 20 speakers (10-male and 10-female utterances of sentences from JNAS database [7]) as the target speech signals and two types of noise signals (white Gaussian noise and railway station noise). By combining the target speech signals and the noise signals, the test data are obtained. The first half of each test data consists of noise only. We assume that noise only signals were estimated completely by voice activity detection. All the signals used in this experiment are 16-kHz-sampled signals. The input SNR of test data was set to 0 dB. The weighting factor α of the decision-directed estimation is 0.97.

In the objective evaluation, we evaluated the performances of noise reduction via three indications. In order to compare the amount of noise reduction and sound quality, we calculate the noise reduction rate (NRR) [8] (output SNR - input SNR in dB) and cepstral distortion (CD) [9] (a measure of the degree of spectral envelope distortion) of processed signals. The number of dimensions of the cepstrum is set to 22 in the evaluation. Furthermore, in the preference test, we evaluated the sound quality of processed

signals subjectively, especially for the human impression of the enhanced speech.

4.2 Experimental Results

First, Fig. 2 shows the shape parameters of true speech signals and those estimated using the proposed algorithm in Sect. 3. Here note that, in order to evaluate the realistic kurtosis of normal utterances, we used the latter half of the test data that include no artificially imposed noise-only part; thus, the estimated kurtosis is just concerned with JNAS sentences. From Fig. 2, we can confirm that the optimal shape parameter of speech p.d.f. can be estimated corresponding to each utterance correctly. Thus, it can be expected that the generalized MMSE STSA estimator can be adapted to various types of signals with different p.d.f. using the shape parameter estimated by the proposed algorithm.

Next, the objective evaluation of noise reduction performance is discussed. Figure 3 shows the results for the average NRR and CD of all the target speakers. In Fig. 3, although the generalized MMSE STSA estimator achieves a larger amount of noise reduction than the original MMSE STSA estimator, it leads to more speech distortion. There-

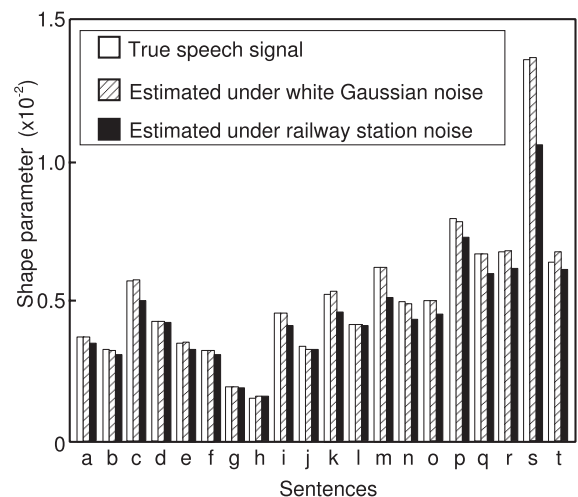


Fig. 2 Estimated shape parameter corresponding to 20 utterances under white Gaussian noise and railway station noise.

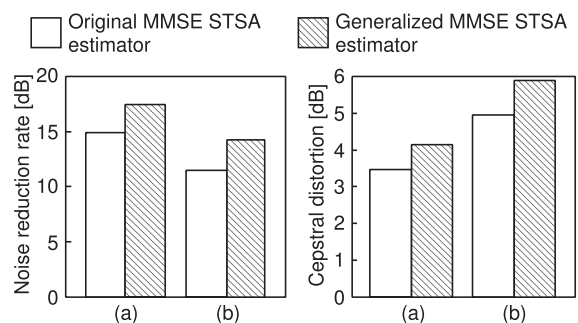


Fig. 3 Results of noise reduction rate and cepstral distortion for (a) white Gaussian noise, (b) railway station noise.

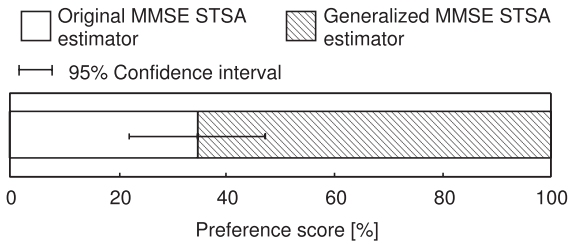


Fig. 4 Result of preference test.

fore, there exists a trade-off between the amount of noise reduction and speech distortion in the original and generalized MMSE STSA estimators.

Finally, the result of a preference test is shown in Fig. 4. Eleven examinees participated in the preference test, in which a pair of processed signals with equivalent NRR using the original MMSE STSA estimator and the generalized MMSE STSA estimator were presented, and the participants were asked to select which signal they preferred. From Fig. 4, the generalized MMSE STSA estimator gains a higher preference score than the original MMSE STSA estimator.

5. Conclusions

In this paper, we proposed a new algorithm to estimate the shape parameter of speech p.d.f. from kurtosis for the generalized MMSE STSA estimator. In the experiment, we confirmed that the generalized MMSE STSA estimator can be adapted to various types of signals with different p.d.f. using the proposed algorithm. According to objective evaluation, it was shown that there is trade-off between the amount of noise reduction and speech distortion at the original and generalized MMSE STSA estimators. However, in the preference test, the generalized MMSE STSA estimator indicated a higher preference score than the original MMSE STSA estimator. Thus, the generalized MMSE STSA estimator with estimation of the shape parameter of speech p.d.f. has an advantage in terms of sound quality over the original MMSE STSA estimator.

Acknowledgements

This work was supported by MIC SCOPE and JST Core Research of Evolutional Science and Technology (CREST), Japan.

References

- [1] P.C. Loizou, *Speech Enhancement Theory and Practice*, CRC Press, Taylor & Francis Group, FL, 2007.
- [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.ASSP-32, no.6 pp.1109–1121, 1984.
- [3] R. Prasad, H. Saruwatari, and K. Shikano, "Estimation of shape pa-

- rameter of GGD function by negentropy matching," *Neural Processing Letters*, vol.22, pp.377–389, 2005.
- [4] E.W. Stacy, "A generalization of the gamma distribution," *The Annals of Math. Statistics.*, vol.33, no.3, pp.1187–1192, 1962.
- [5] I. Andrianakis and P.R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," *Proc. ICASSP*, vol.1071, pp.III-1068–III-1071, 2006.
- [6] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," *Proc. ICASSP*, pp.4037–4040, 2008.
- [7] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn. (E)*, vol.21, no.3, pp.199–206, 1999.
- [8] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing.*, vol.2003, no.11, pp.1134–1146, 2003.
- [9] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice Hall PTR, Upper Saddle River, NJ, 1993.

Appendix: Fitness of Statistical Model for Speech

To show justification of using the statistical speech model, we evaluated fitness of the generalized gamma distribution for the speech amplitude spectrum. We compared shapes of a histogram obtained from the real speech signal's amplitude spectrum and p.d.f. of the generalized gamma distribution with the optimal shape parameter estimated from the subjective speech signal.

We applied Kolmogorov-Smirnov (KS) test to 20 JNAS utterances used in Sect. 4. As the result, we can confirm that p.d.f.s. of the optimized generalized gamma distributions fit each of 15 utterances out of 20 utterances at the significance level of 5%. Figure A-1 shows an example of the real histogram of the typical speech amplitude spectrum and corresponding p.d.f. of the generalized gamma distribution with the optimal shape parameter. This result well indicates that the use of the generalized gamma distribution can be justified to some extent.

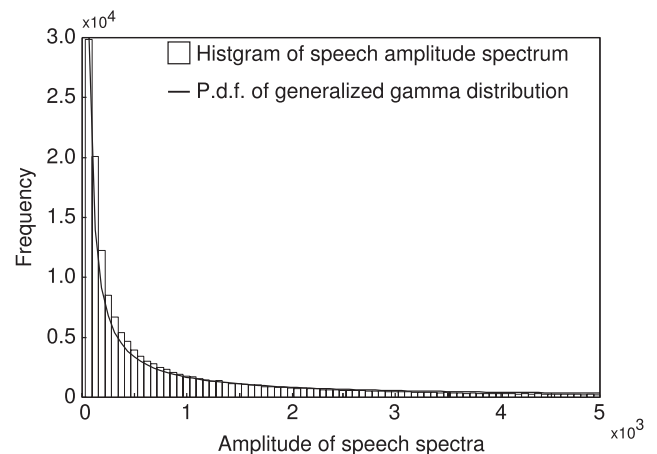


Fig. A-1 Example of speech spectral amplitude histogram and corresponding p.d.f. of generalized gamma distribution.