STATISTICAL MODELS FOR SPEECH DEREVERBERATION

Takuya Yoshioka[1,2], *Hirokazu Kameoka*[1], *Tomohiro Nakatani*[1], *and Hiroshi G. Okuno*[2]

[1]NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[2]Graduate School of Informatics, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto 606-8501, Japan
email: takuya@cslab.kecl.ntt.co.jp, web: http://www.kecl.ntt.co.jp/icl/signal/takuya/index.html

ABSTRACT

This paper discusses a statistical-model-based approach to speech dereverberation. With this approach, we first define parametric statistical models of probability density functions (pdfs) for a clean speech signal and a room transmission channel, then estimate the model parameters, and finally recover the clean speech signal by using the pdfs with the estimated parameter values. The key to the success of this approach lies in the definition of the models of the clean speech signal and room transmission channel pdfs. This paper presents several statistical models (including newly proposed ones) and compares them in a large-scale experiment. As regards the room transmission channel pdf, an autoregressive (AR) model, an autoregressive power spectral density (ARPSD) model, and a moving-average power spectral density (MAPSD) model are considered. A clean speech signal pdf model is selected according to the room transmission channel pdf model. The AR model exhibited the highest dereverberation accuracy when a reverberant speech signal of 2 sec or longer was available while the other two models outperformed the AR model when only a 1-sec reverberant speech signal was available.

Index Terms— Dereverberation, statistical model.

1. MOTIVATION

Reverberation degrades the quality of speech picked up by distant microphones and thereby limits the applicability of existing speech processing products. Therefore, dereverberation techniques, which mitigate the unfavorable reverberation effect, are vital for the further expansion of existing speech processing products as well as for the development of new ones. In fact, many dereverberation methods have been proposed including a maximum kurtosis method [1], a spectral subtraction method [2], and a weighted prediction error (WPE) method [3].

In practical situations, speech quality may be degraded not only by reverberation but also by ambient noise and interfering speech. Therefore, dereverberation techniques should be designed so that they can be effectively combined with noise reduction and speech separation techniques.

A statistical-model-based approach has been employed extensively for noise reduction and speech separation [4]. Thus, this approach affords a good guide to combining the dereverberation techniques with noise reduction and speech separation techniques. In fact, on the basis of the WPE method for dereverberation [3], which is based on this approach, we have developed several dereverberation methods coupled with such other signal processing techniques (e.g., [5]). This paper further investigates the statistical-model-based approach to dereverberation.

Now, we briefly review the concept of the statistical-model-based approach to dereverberation. We begin by formulating the dereverberation problem. In this paper, we deal with a single microphone case. Let $s_{n,l}$ and $y_{n,l}$ denote clean and reverberant speech signals, respectively, represented in the short-time Fourier transform (STFT) domain, where n and l are time frame and frequency bin indices, respectively. Assume that we observe $y_{n,l}$ over N consecutive time frames. We represent the observed data set as $y = \{y_{n,l}\}_{0\leq n<N,0\leq l<L}$, where L is the number of frequency bins. In response to this, we denote the corresponding set of clean speech samples by $s = \{s_{n,l}\}_{0\leq n<N,0\leq l<L}$. Dereverberation is a process for estimating s when y is given.

With the statistical-model-based approach, we first define statistical models of a room transmission channel and a clean speech signal by using probability density functions (pdfs) $p(y|s,\Theta)$ and $p(s|\Theta)$, respectively, where Θ is the set of all parameters. We refer to the respective pdfs as the room acoustics pdf and the clean speech pdf. Then, we estimate Θ values using the observed data, y. We denote by $\hat{\Theta}$ the set of the estimated parameter values. Finally, we compute the minimum mean square error (MMSE) estimate of the clean speech signal as

$$\hat{s} = E_s\{s; p(s|y,\hat{\Theta})\}, \tag{1}$$

where $p(s|y,\hat{\Theta}) \propto p(y|s,\hat{\Theta})p(s|\hat{\Theta})$ and $E_x\{f(x);q(x)\} = \int q(x)f(x)dx$. In summary, we need to perform the following three steps to derive a specific dereverberation method.

(s1) Define a room acoustics pdf, $p(y|s,\Theta)$, and a clean speech pdf, $p(s|\Theta)$, along with a parameter set, Θ.

(s2) Define an estimation algorithm for Θ.

(s3) Derive the MMSE signal estimator given by (1).

The key to the successful design of a dereverberation method is the selection of statistical models of room acoustics and clean speech pdfs. Different models lead to dereverberation methods with different characteristics. Therefore, we must understand what kind of dereverberation method is derived based on each model.

With the above motivation, this paper describes several statistical models of room acoustics and clean speech pdfs, and experimentally compares the dereverberation methods derived from these models. The models considered in this paper include both newly proposed and already reported models [3]. The comparison is carried out in terms of maximum dereverberation accuary and

978-1-4244-3679-8/09/$25.00 ©2009 IEEE

the amount of observation data required for model parameter estimation. (Although computational cost and memory size are also important, we do not cover these performance indices because they are dependent on implementation.)

2. MODELS CONSIDERED IN THIS PAPER

In Section 2, we describe several statistical models of room acoustics and clean speech pdfs.

2.1. Basic PDFs

First, we present two basic pdfs that are used in the definition of room acoustics and clean speech pdfs.

- *Complex normal distribution*
  $\mathcal{N}_{\mathbb{C}}\{x;\mu,\sigma^2\}$ denotes the pdf of a complex normal distribution with mean μ and variance σ^2, i.e.,

$$\mathcal{N}_{\mathbb{C}}\{x;\mu,\sigma^2\} = \frac{1}{\pi\sigma^2}\exp\Big\{\frac{|x-\mu|^2}{\sigma^2}\Big\} \quad \text{for } x \in \mathbb{C}. \tag{2}$$

- *Generalized gamma distribution*
  $\mathcal{GG}\{x;\kappa,\theta,\rho\}$ denotes the pdf of a generalized gamma distribution with scale parameter θ and two shape parameters κ and ρ, i.e.,

$$\mathcal{GG}\{x;\kappa,\theta,\rho\} = \frac{\rho x^{\kappa\rho-1}}{\Gamma(\kappa)\theta^{\kappa\rho}}\exp\Big\{-\Big(\frac{x}{\theta}\Big)^{\rho}\Big\} \quad \text{for } x > 0. \tag{3}$$

2.2. Room acoustics PDFs

Now, we describe the room acoustics pdfs, $p(y|s,\Theta)$, that are considered in this paper. We assume that reverberant speech signals at different frequency bins are statistically independent of each other. Then, $p(y|s,\Theta)$ can be decomposed as

$$p(y|s,\Theta) = \prod_{l=0}^{L-1}\prod_{n=0}^{N-1} p(y_{n,l}|y_0^{n-1},s,\Theta), \tag{4}$$

where $y_0^{n-1} = \{y_{n',l}\}_{0\le n'<n, 0\le l<L}$. The pdf on the right hand side of (4), $p(y_{n,l}|y_0^{n-1},s,\Theta)$, represents the probability of observing $y_{n,l}$ at time frame n on condition that we have the clean speech signal and the past observed signal. In this paper, we consider three statistical models for this conditional pdf. Two of these three models are presented for the first time.

Autoregressive (AR) model

The first room acoustics model is an autoregressive (AR) model, which we described in our previous paper [3]. (Note however that, in [3], we did not restrict the number of microphones to one.) The reverberant speech signal, $y_{n,l}$, is divided into its clean component, $s_{n,l}$, and reverberation component, denoted by $r_{n,l}$, as

$$y_{n,l} = s_{n,l} + r_{n,l}. \tag{5}$$

The AR model assumes that a room transmission channel is modeled by an AR system, or equivalently, that $r_{n,l}$ is given by

$$r_{n,l} = \sum_{k=D_l}^{D_l+K_l-1} h_{k,l}^{*} y_{n-k,l}, \tag{6}$$

where $h_{k,l}$ is a complex number and $*$ is a complex conjugate. Therefore, the conditional pdf, $p(y_{n,l}|y_0^{n-1},s,\Theta)$, is described as

$$p(y_{n,l}|y_0^{n-1},s,\Theta) = \mathcal{N}_{\mathbb{C}}\Big\{y_{n,l};\sum_{k=D_l}^{D_l+K_l-1} h_{k,l}^{*} y_{n-k,l} + s_{n,l},\sigma^2\Big\} \quad (\sigma^2\to 0). \tag{7}$$

By substituting (7) into (4), we obtain the AR-model-based room acoustics pdf. This model is parameterized by $\{\{h_{k,l}\}_{D_l\le k<D_l+K_l}\}_{0\le l<L}$; in other words, we have $h_{k,l}\in\Theta$. The validity of the AR model is discussed in [3].

Reverberation generally affects both the power and phase spectra of a clean speech signal. We see that (7) takes account of the reverberation effects on both the power and phase spectra. By contrast, the remaining two room acoustics models consider only the reverberation effect on the power spectra.

Autoregressive psd (ARPSD) model

The second room acoustics model is an autoregressive power spectral density (ARPSD) model. This model is new and inspired by the Lebart et al.'s work [2]. Lebart et al. assumed that the power spectral density (psd) of the reverberation component of a reverberant speech signal is the same as the psd of a delayed version of the reverberant speech signal up to a constant scale factor. This assumption was derived from the observation that sound energy exponentially decays in a room. (Later, Habets et al. reported that the exactness of this assumption can be improved by further taking into consideration direct-to-reverberation ratio [6].) The delay amount is set at about 50 msec, and the constant scale factor is determined with respect to the reverberation time.

The ARPSD model regards the reverberation component, $r_{n,l}$, as an additive noise. In other words, we assume that $r_{n,l}$ is uncorrelated with $s_{n,l}$. Furthermore, generalizing the assumption made by Lebart et al., we assume that $r_{n,l}$ is normally distributed with a mean of 0 and a variance (i.e., a psd component) given by an autoregressive form as

$$p(r_{n,l}|\Theta) = \mathcal{N}_{\mathbb{C}}\Big\{r_{n,l};0,\Big(\sum_{k=D_l}^{D_l+K_l-1} g_{k,l}|y_{n-k,l}|^2\Big)\Big\}. \tag{8}$$

Note that the $g_{k,l}$ value is non-negative. Based on (8), we obtain the following conditional pdf:

$$p(y_{n,l}|y_0^{n-1},s,\Theta) = \mathcal{N}_{\mathbb{C}}\Big\{y_{n,l};s_{n,l},\Big(\sum_{k=D_l}^{D_l+K_l-1} g_{k,l}|y_{n-k,l}|^2\Big)\Big\}. \tag{9}$$

If we set K_l at 1 and D_l at about 50 msec, (9) becomes equivalent to the model of [2]. The ARPSD model is parameterized by $\{\{g_{k,l}\}_{D_l\le k<D_l+K_l}\}_{0\le l<L}$; thus we have $g_{k,l}\in\Theta$.

Incidentally, the method described in [2] determines the $g_{k,l}$ value based on the reverberation time. Lebart et al. proposed estimating the broadband reverberation time by exploiting short pauses in speech activity. However, because a reverberation time generally depends on a frequency bin, we may not achieve an acceptable dereverberation accuracy by using only the broadband reverberation time. In constrast, we can estimate $g_{k,l}$ without using the reverberation time by combining the ARPSD model with a sparseness-constrained non-stationary Gaussian model of a clean speech pdf, which we describe later.

Moving-average psd (MAPSD) model

The third model is a moving-average psd (MAPSD) model. This model also regards reverberation component $r_{n,l}$ as an additive noise. The difference from the ARPSD model lies in the assumption that the variance (i.e., the psd component) of $r_{n,l}$ is expressed by using a moving-average (MA) form. Specifically, the conditional pdf on the right hand side of (4) is assumed to be given by

$$p(y_{n,l}|y_0^{n-1}, s, \Theta) = \mathcal{N}_{\mathbb{C}}\Big\{y_{n,l}; s_{n,l}, \sum_{k=D_l}^{D_l+K_l-1} \frac{g_{k,l}}{\eta_{n-k,l}}\Big\}, \quad (10)$$

where $\eta_{n,l}$ is the reciprocal of the psd component, which is sometimes called the accuracy, of clean speech signal $s_{n,l}$. This model is parameterized by $\{\{g_{k,l}\}_{D_l \le k < D_l+K_l}\}_{0 \le l < L}$; thus, we have $g_{k,l} \in \Theta$. The MAPSD model aims to remove the exponential decay assumption made by the ARPSD model.

The above three room acoustics models are compared in Section 4.

2.3. Clean speech PDFs

Next, let us describe the clean speech models. We start with the assumption that $s_{n,l}$ and $s_{n',l'}$ are statistically independent if $(n,l) \neq (n',l')$. Then, we obtain

$$p(s|\Theta) = \prod_{l=0}^{L-1} \prod_{n=0}^{N-1} p(s_{n,l}|\Theta). \quad (11)$$

As regards $p(s_{n,l}|\Theta)$, we consider the following two models:

- an unconstrained non-stationary Gaussain (UCNSG) model;
- a sparseness-constrained non-stationary Gaussian (SCNSG) model.

A clean speech pdf model should be selected in accordance with a room acoustics model so that we can derive an effective parameter estimator in step (s2). The UCNSG model is used in combination with the AR model. On the other hand, the SCNSG model is used with the ARPSD or MAPSD models.

Unconstrained non-stationary Gaussian (UCNSG) model

In the field of independent component analysis, it is well known that effective speech separation algorithms can be obtained by assuming a clean speech signal to be a realization of a non-stationary Gaussian process. With this as a basis, we consider the following model of $p(s_{n,l}|\Theta)$:

$$p(s_{n,l}|\Theta) = \mathcal{N}_{\mathbb{C}}\{s_{n,l}; 0, 1/\eta_{n,l}\}. \quad (12)$$

The parameter set for this model is the set of accuracies, $\{\eta_{n,l}\}_{0 \le n < N, 0 \le l < L}$; thus, we have $\eta_{n,l} \in \Theta$.

Sparseness-constrained non-stationary Gaussian (SCNSG) model

A clean speech signal generally contains short pauses and has a harmonic spectral structure. Therefore, the psd components of a clean speech signal are sparsely distributed. To express this sparseness, the second model uses the following prior pdf for accuracy $\eta_{n,l}$ in addition to (12):

$$p(\eta_{n,l}) = \mathcal{GG}\{\eta_{n,l}; \kappa, \theta, \rho\}, \quad (13)$$

where κ, θ, ρ are prescribed constants.

In summary, we consider the following three combinations of room acoustics and clean speech pdfs:

(m1) an AR model combined with an UCNSG model;

(m2) an ARPSD model combined with an SCNSG model;

(m3) an MAPSD model combined with an SCNSG model.

Thus, step (s1) has been completed.

3. DEREVERBERATION METHODS

By carrying out steps (s2) and (s3) for each statistical model combination described in Section 2, we can derive dereverberation methods corresponding to the respective combinations. Combination (m1) leads to the WPE method [3]. In this section, we describe the parameter estimation algorithm (step (s2)) and MMSE signal estimator (step (s3)) for combination (m2). The parameter estimation algorithm and MMSE signal estimator for combination (m3) can be derived in a similar way to those for (m2).

The statistical models for (m2) are defined by (4), (9), (11), (12), and (13). Hence, the set of parameters is given by

$$\Theta = \{\{g_{k,l}\}_{D_l \le k < D_l+K_l}, \{\eta_{n,l}\}_{0 \le n < N}\}_{0 \le l < L}. \quad (14)$$

We estimate these parameter values by means of maximum a posteriori (MAP) estimation, i.e., we compute $\hat{\Theta}$ that maximizes the following posterior pdf:

$$p(\Theta|y) \propto p(\Theta) E_s\{p(y|s,\Theta); p(s|\Theta)\}. \quad (15)$$

The first term on the right hand side of (15) is defined as $p(\Theta) = \prod_{n=0}^{N-1} \prod_{l=0}^{L-1} p(\eta_{n,l})$, where we ignore the prior for $g_{k,l}$. $p(y|s,\Theta)$ is given by (4) and (9), while $p(s|\Theta)$ is given by (11) and (12).

$\hat{\Theta}$ that maximizes the posterior pdf, (15), cannot be analytically calculated. There are two factors for this difficulty. One is that each variance of the ARPSD model, (9), is defined as a (weighted) sum of $g_{k,l}$ over several k values. The other is that the log of the prior, given by (13), involves the computation of ρ power of $\eta_{n,l}$. To cope with the first problem, we use an expectation-maximization (EM) algorithm by introducing latent variables $r_{D_l,n,l}, \cdots, r_{D_l+K_l-1,n,l}$ for each n and l. $r_{k,n,l}$ represents the component of reverberation component $r_{n,l}$ that comes from the reverberant speech signal at the $(n-k)$th time frame, $y_{n-k,l}$. Therefore, $r_{k,n,l}$ satisfies the following two conditions.

$$r_{D_l,n,l} + \cdots + r_{D_l+K_l-1,n,l} = r_{n,l} \quad (16)$$

$$p(r_{k,n,l}|\Theta) = \mathcal{N}_{\mathbb{C}}\{r_{k,n,l}; 0, g_{k,l}|y_{n-k,l}|^2\}. \quad (17)$$

We cope with the latter problem by using an auxiliary function method, which is commonly used in the field of non-negative matrix factorization (NMF) [7].

The parameter estimation algorithm may be described as follows (we omit the detailed derivation owing to the limited space). In the following, we use vector $\boldsymbol{r}_{n,l}$, which is defined as $\boldsymbol{r}_{n,l} = [r_{D_l,n,l}, \cdots, r_{D_l+K_l-1,n,l}]^T$. First , we set initial parameter values, denoted by $\Theta^{[0]}$. Then, we repeat the following two steps until convergence.

1. For all n and l values, compute $p(\boldsymbol{r}_{n,l}|y,\Theta^{[i]})$ according to the following equations (E-step):

$$p(\boldsymbol{r}_{n,l}|y,\Theta^{[i]}) = \mathcal{N}_{\mathbb{C}}\{\boldsymbol{r}_{n,l};\boldsymbol{\mu}_{n,l}^{[i]},\Sigma_{n,l}^{[i]}\} \quad (18)$$

$$\boldsymbol{\mu}_{n,l}^{[i]} = (W_{n,l}^{[i]})^H \mathbf{1} y_{n,l} \quad (19)$$

$$\Sigma_{n,l}^{[i]} = (\Lambda_{n,l}^{[i]}/\eta_{n,l}^{[i]})(I/\eta_{n,l}^{[i]} + \mathbf{1}\mathbf{1}^T\Lambda_{n,l}^{[i]})^{-1} \quad (20)$$

$$W_{n,l}^{[i]} = \Lambda_{n,l}^{[i]}(I/\eta_{n,l}^{[i]} + \mathbf{1}\mathbf{1}^T\Lambda_{n,l}^{[i]})^{-1}. \quad (21)$$

I and $\mathbf{1}$ denote the K_l-dimensional identity matrix and all-one vector, respectively, and $\Lambda_{n,l}^{[i]}$ is given as follows:

$$\Lambda_{n,l}^{[i]} = \mathrm{diag}(g_{D_l,l}^{[i]}|y_{n-D_l,l}|^2, \cdots, g_{D_l+K_l-1,l}^{[i]}|y_{n-D_l-K_l+1,l}|^2). \quad (22)$$

Note that the matrix inversions in (20) and (21) can be avoided by using the Woodbury fomula.

2. For all n, l, and k values, update the parameter values according to the following equations (M-step):

$$1/\eta_{n,l}^{[i+1]} = E_{\boldsymbol{r}_{n,l}}\{|y_{n,l} - \boldsymbol{g}_l^T\boldsymbol{r}_{n,l}|^2; p(\boldsymbol{r}_{n,l}|y,\Theta^{[i]})\}/\kappa\rho + 1/\kappa\theta^\rho(\eta_{n,l}^{[i]})^{1-\rho} \quad (23)$$

$$g_{k,l}^{[i+1]} = \sum_{n=0}^{N-1} \frac{E_{r_{k,n,l}}\{|r_{k,n,l}|^2; p(\boldsymbol{r}_{n,l};y,\Theta^{[i]})\}}{N|y_{n-k,l}|^2}. \quad (24)$$

Because $p(\boldsymbol{r}_{n,l}|y,\Theta^{[i]})$ is based on a complex normal distribution as shown by (18), the expectated values in (23) and (24) can be readily computed.

As regards step (s3), it is obvious that the MMSE signal estimator is given by a Wiener filter. This is a consequence of the fact that we regarded the reverberation component, $r_{n,l}$, as an additive noise. We omit the details due to lack of space.

4. EXPERIMENT AND CONCLUSION

We conducted an experiment to compare the dereverberation methods derived from the statistical models described in Section 2. We used utterances of 306 speakers contained in the JNAS database. For each speaker, we made 1-, 2-, 3-, 4-, and 5-sec clean speech signals with a sampling rate of 16 kHz. The individual clean speech signals were convolved with an impulse response measured in a room with a reverberation time of 0.6 sec to simulate reverberant speech signals. The frame size and frame shift for STFT were set at 512 points (32 msec) and 128 points (8 msec), respectively. D_l, K_l, κ, ρ, and θ values were experimentally determined. The experimental results were evaluated in terms of the 12th-order cepstral distances (CD) between the target (i.e., processed or reverberant) speech signals and the corresponding clean speech signals. The CD was shown to be highly correlated with subjective quality of dereverberated speech [8].

Figure 1 shows the CDs averaged over the 306 speakers against the amount of observed data. We see that combination (m1) provided the smallest average CDs when reverberant speech signals of 2 sec or longer were available. However, with (m1), the average CD increased when we used only 1-sec reverberant speech signals. By contrast, we find that combinations (m2) and (m3) constantly improved the average CDs to some degree. Although the speech signals processed by the dereverberation methods for (m2) and (m3) sounded somewhat distorted due to Wiener filtering, they were much less reverberant than the input speech signals.

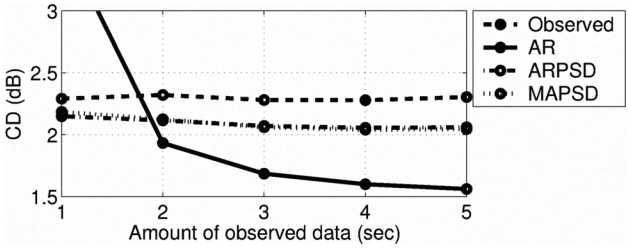

Figure 1: Average cepstral distances. Note that the lines for ARPSD and MSPSD models are almost completely overlapping.

From the above results, we conclude that, for a room acoustics pdf, we should use an AR model when we can assume that a speaker does not move frequently. For situations where this assumption is invalid, ARPSD and MAPSD models are better. Effectively combining these two models would achieve dereverberation with high accuracy using only a small amount of observation data.

In this paper, we focused on maximum dereverberation accuracy and observation data amount. Another important performance index is sensitivity to environmental changes. We will conduct comparative tests in relation to these other performance indices.

5. REFERENCES

[1] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, vol. VI, 2001, pp. 3701–3704.

[2] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, 2001.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. Int'l Conf. Acoust. Speech, Signal Process.*, 2008, pp. 85–88.

[4] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer, 2005.

[5] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, 2009.

[6] E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *Proc. Int'. Symp. Signal Process., Inf. Tech.*, 2006, pp. 651–655.

[7] Y. Lin and D. D. Lee, "Bayesian L1-norm sparse learning," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, vol. V, 2006, pp. 605–608.

[8] K. Furuya and A. Kataoka, "Factor analysis of speech quality improved by dereverberation processing," *IEICE Trans. Fund.*, vol. E91-A, no. 8, pp. 763–771, 2008, in Japanese.