A Countermeasure Method For Background Noise By jointly Using Bone- and Air- Conducted Speech Signals.

3 authors:

Gerard Gallagher
Inria Rennes - Bretagne Atlantique

Hisako Orimoto
Prefectural University of Hiroshima

Akira Ikuta
Prefectural University of Hiroshima

25th International Congress on Sound and Vibration
8-12 July 2018 HIROSHIMA CALLING
ICSV25

A COUNTERMEASURE METHOD FOR BACKGROUND NOISE BY JOINTLY USING BONE- AND AIR- CONDUCTED SPEECH SIGNALS

Gerard Gallagher, Akira Ikuta and Hisako Orimoto

Prefectural University of Hiroshima, Faculty of Management and Information Systems, Hiroshima, Japan

email:q696001hz@ed.pu-hiroshima.ac.jp, {ikuta,orimoto}@pu-hiroshima.ac.jp

When applying speech recognition to actual circumstances such as inspection and maintenance operations in industrial factories to recording and reporting routines at construction sites, etc. where hand-writing is difficult, some countermeasure methods for surrounding noise are indispensable. In this study, a signal processing method to remove the noise for actual speech signals is proposed by jointly using the measured data of bone- and air- conducted speech signals. More specifically, by introducing Bayes' theorem based on an orthogonal expansion expression of probability distribution, a new type of noise removal method is proposed. The effectiveness of the proposed method is confirmed by applying it to bone- and air- conducted speech measured in real environments under the existence of surrounding noise.

Keywords: Noise suppression, Speech signals

1. Introduction

Speech recognition systems have been applied in various fields due to recent development of digital signal processing techniques. For example, these systems are applied to inspection and maintenance operations in industrial factories and to recording and reporting routines at construction sites, etc. where hand-writing is difficult. For speech recognition in such actual circumstances, some countermeasure methods for surrounding noises are indispensable. Several noise suppression methods for the speech signal have been proposed up to now.

Previously reported methods for noise reduction in speech recognition can be classified into two categories. One is based on a single microphone [1],[2] and the other uses a microphone array [3]. Since the latter requires priori information on the number of noise sources, and the number of microphones needed is larger than that of the noise sources in the case of multi-noise sources, this category demands large scale systems. Therefore, the former based on a single microphone is more advantageous than the latter [4],[5].

In such a noise suppression task for speech signals based on a single microphone, many algorithms applying the Kalman filter have been proposed up to now [6]-[8]. However, the Kalman filter is originally based on the assumption of Gaussian white noise [9],[10]. However, the actual noises show complex fluctuation forms with non-Gaussian and non-white properties.

From the above viewpoint of the assumption of Gaussian white noise has been proposed [11]. The method can be applied to actual complex situations where both the noise statistics and the fluctuation forms of speech , in our previously reported study, a noise suppression algorithm for the actual speech signals without requirement, signals are unknown. Furthermore, by applying the algorithm to real speech signals with several kinds of noises, its effectiveness has been experimentally confirmed in comparison with the Kalman filter.

In this study, a signal processing method to remove noise for actual speech signal is proposed by jointly using the measured data of bone- and air- conducted speech signals. Though the bone-conducted speech signal is a kind of solid propagation sound with less effect by the surrounding noise, the high frequency component of the signal reduces through the propagation process. On the other hand, air-conducted speech signal contains all frequency components though the signal is strongly affected by the surrounding noise. Therefore, by using jointly both bone- and air- conducted speech signals, a more accurate estimation of the speech signal can be expected whilst recovering the high-frequency components of the speech signal even in a very noisy circumstance.

By introducing Bayes' theorem based on an orthogonal expansion expression of probability distribution, a new type of noise removal method is proposed. Using the principal of the Unscented Kalman filter a deterministic minimal set of sample points around the mean called sigma points [12] are introduced in order to simplify the calculation process of the coefficients. The effectiveness of the proposed method is confirmed by applying it to bone- and air-conducted speech measured in a real environment under the existence of surrounding noise.

2. NOISE SUPPRESSION METHOD FOR SPEECH SIGNALS

2.1 Stochastic Model for Air- and Bone- Conducted Speech Signals

In the actual environment with a surrounding noise, let x_k, y_k, and y_k, be the original speech signal, the observations of air- and bone- conducted speech signal at a discrete time k. The observation y_k, is contaminated by a background noise v_k. According to the additive property of sound pressure, the following relationship can be established.

$$y_k = x_k + v_k, \tag{1}$$

Where the statistics of v_k are assumed to be known.

In order to estimate the parameters of the propagation model for bone-conducted speech signals, the correlation information between speech signal x_k and observed signal z_k is generally necessary. However, it is difficult to find the information in advance because x_k is an unknown signal to be estimated. In this study, a new adaptive algorithm for noise suppression is proposed by introducing a propagation model with unknown parameters between x_k and z_k as the bone-conducted speech signal models for z_k:

$$z_k = a_k x_k + b_k w_k \tag{2}$$

Where w_k is a random noise with zero-mean and variance 1, and a_k and b_k are unknown parameters. In order to estimate the parameters simultaneously with the speech signal x_k, the following dynamic models of the parameters are introduced.

$$a_{k+1=}a_k,\ \ b_{k+1=}b_k. \tag{3}$$

2.2 Derivation of an Adaptive Noise Suppression Algorithm Based on Bayes' Theorem

To derive an estimation algorithm for the speech signal x_k we place our basis on Bayes' theorem for the conditional probability distribution [13],[14]. Since the parameter a_k and b_k are also unknown, the conditional probability distribution of x_k, a_k and $b_k$is expressed by

$$P(x_k, a_k, b_k | Y_k, Z_k) = \frac{P(x_k, a_k, b_k, y_k, z_k | Y_{k-1}, Z_{k-1})}{P(y_k | Y_{k-1}, Z_{k-1})} \tag{4}$$

where$Y_k (= \{y_1, y_2, \ldots., y_k\})$and $Z_k (= \{z_1, z_2, \ldots., z_k\})$are sets of air- and bone- conducted speech signal data up to time k. By expanding the conditional joint probability distribution $P(x_k, a_k, b_{k,} y_k, z_k | Y_{k-1}, Z_{k-1})$ in a statistical orthogonal expansion series on the basis of well-known standard probability distributions, which describe the dominant part of the actual fluctuation, the following expression is derived.

$$P(x_k, a_k, b_k | Y_k, Z_k)$$

$$= \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} A_{lmnpq}\, P_0(x_k | Y_{k-1}, Z_{k-1})\, P_0(a_k | Y_{k-1}, Z_{k-1}) P_0(b_k | Y_{k-1}, Z_{k-1})$$

$$\cdot\, \varphi_l^{(1)}(x_k) \varphi_m^{(2)}(a_k)\, \varphi_n^{(3)}(b_k) \varphi_p^{(4)}(y_k) \varphi_q^{(5)}(z_k) / \textstyle\sum_{p=0}^{\infty} \sum_{q=0}^{\infty} A_{000pq} \varphi_p^{(4)}(y_k) \varphi_q^{(5)}(z_k) \tag{5}$$

with

$$A_{lmnpq} \equiv < \varphi_l^{(1)}(x_k) \varphi_m^{(2)}(a_k)\, \varphi_n^{(3)}(b_k) \varphi_p^{(4)}(y_k) \varphi_q^{(5)}(z_k) | Y_{k-1}, Z_{k-1} > \tag{6}$$

The above four functions $\varphi_l^{(1)}(x_k), \varphi_m^{(2)}(a_k), \varphi_n^{(3)}(b_k), \varphi_p^{(4)}(y_k)$ and $\varphi_q^{(5)}(z_k)$ are orthonormal polynomials of degree $l, m, n, p\ and\ q$ with weighting functions,

$P_0(x_k | Y_{k-1}, Z_{k-1})$, $P_0(a_k | Y_{k-1}, Z_{k-1})$, $P_0(y_k | Y_{k-1}, Z_{k-1})$ and $P_0(z_k | Y_{k-1}, Z_{k-1})$.

As examples of standard probability functions for the speech signal, the parameters and observations of the air- and bone- conducted speech signal, we adopt Gaussian distributions with means and variances, as

$$
\begin{aligned}
x_k^* &\equiv < x_k | Y_{k-1}, Z_{k-1} >, & \Gamma x_k &\equiv < (x_k - x_k^*)^2 | Y_{k-1}, Z_{k-1} >, \\
a_k^* &\equiv < a_k | Y_{k-1}, Z_{k-1} >, & \Gamma a_k &\equiv < (a_k - a_k^*)^2 | Y_{k-1}, Z_{k-1} >, \\
b_k^* &\equiv < b_k | Y_{k-1}, Z_{k-1} >, & \Gamma b_k &\equiv < (b_k - b_k^*)^2 | Y_{k-1}, Z_{k-1} >, \\
y_k^* &\equiv < y_k | Y_{k-1}, Z_{k-1} >, & \Omega_k &\equiv < (y_k - y_k^*)^2 | Y_{k-1}, Z_{k-1} >, \\
z_k^* &\equiv < z_k | Y_{k-1}, Z_{k-1} >, & \Phi_k &\equiv < (z_k - z_k^*)^2 | Y_{k-1}, Z_{k-1} >.
\end{aligned} \tag{7}
$$

Sigma points are introduced to both calculate the optimal terms and reduce the complexity of the calculation process. The sigma points are calculated at 3 different points in a recursive fashion. The sigma points are calculated with calculated for the following variables x_k, a_k, b_k, y_k, z_k with W being the weighting to be used. The following is the equation for the sigma point of variable x_k.

$$
\begin{aligned}
\sigma x(1) &= x_k^* \\
\sigma x(2) &= x_k^* + \sqrt{1 + \lambda * \Gamma x_k} \\
\sigma x(3) &= x_k^* - \sqrt{1 + \lambda * \Gamma x_k}
\end{aligned} \tag{8}
$$

The non-Gaussian properties of the speech signal and observations of the air- and bone conducted speech signal are reflected in each expansion coefficient A_{lmnpq}. The orthonormal polynomials with the weighting probability distributions are then specified as a Hermite polynomial [15].

Based on (5), the estimate of the polynomial function $f_{L,M,N}(x_k, a_k, b_k)$ of x_k, a_k and b_k with $(L,M,N)^{\text{th}}$ order can be derived as follows.

$$
\begin{aligned}
\hat{f}_{L,M,N}(x_k, a_k, b_k) &\equiv < f_{L,M,N}(x_k, a_k, b_k) | Y_k Z_k > \\
&= \frac{\sum_{l=0}^{L}\sum_{m=0}^{M}\sum_{n=0}^{N}\sum_{p=0}^{\infty}\sum_{q=0}^{L} C_{lmn}^{LMN} A_{lmnpq} \varphi_p^{(4)}(y_k)\varphi_q^{(5)}(z_k)}{\sum_{p=0}^{\infty}\sum_{q=0}^{\infty} A_{000pq}\varphi_p^{(4)}(y_k)\varphi_q^{(5)}(z_k)},
\end{aligned} \tag{9}
$$

where C_{lmn}^{LMN} is an appropriate constant satisfying the following equality:

$$
\hat{f}_{L,M,N}(x_k, a_k, b_k) = \sum_{l=0}^{L}\sum_{m=0}^{M}\sum_{n=0}^{N} C_{lmn}^{LMN}\, \varphi_l^{(1)}(x_k), \varphi_m^{(2)}(a_k), \varphi_n^{(3)}(b_k). \tag{10}
$$

Using the property of conditional expectation, (1) and (2), the variables for the air- and bone-conduced speech signal in (7) can be calculated as follows:

$$
\begin{aligned}
y_k^* &= < x_k + v_k | Y_{k-1}, Z_{k-1} > \\
&= x_k^* + < v_k >
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
\Omega_k &= < (x_k + v_k - x_k^* - < v_k >)^2 | Y_{k-1}, Z_{k-1} > \\
&= \Gamma x_k + < (v_k - < v_k >)^2 >
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
z_k^* &= < a_k x_k + b_k w_k | Y_{k-1}, Z_{k-1} > \\
&= a_k^* x_k^*
\end{aligned} \tag{13}
$$

$$\Phi_k = < (a_k x_k + b_k w_k - a_k^* x_k^*)^2 | Y_{k-1} Z_{k-1} >$$
$$= \left(\Gamma a_k + a_k^{*^2}\right)\left(\Gamma x_k + x_k^{*^2}\right) - a_k^* x_k^* + \Gamma b_k + b_k^{*^2} \quad (14)$$

Furthermore, substituting (1), (2) and (8) into (6), the expansion coefficients A_{lmnpq} are given in functional forms of $x_k^*, \Gamma x_k a_k^*, \Gamma a_k, b_k^*, \Gamma b_k$.

Finally, the following time transition model for the speech signal is generally established.

$$x_{k+1} = F x_k + G u_k \quad (15)$$

Where u_k is a random input with known statistics. Two parameters F and G can be determined on the basis of correlation information for the time series of x_k. By considering (15) and the dynamical models of a_k and b_k in (3), the predictions x_{k+1}^*, Γ_{x+1}, a_{k+1}^*, Γ_{a+1}, b_{k+1}^*, Γ_{b+1} may be then given in functional forms on estimations of x_k, a_k and b_k. Therefore, the estimation of the speech signal can be performed in a recursive way.

3. APPLICATION TO SPEECH SIGNALS IN REAL ENVIRONMENTS

In order to confirm the effectiveness of the proposed noise suppression algorithm, it was applied to real speech signals. The speech signal data was measured in the Prefectural University of Hiroshima. The speech signals language is English, consisting of 15 syllables. More specifically, for a female and a male speech signals digitized with sampling frequency of 10kHz and quantization of 16 bits, we estimated the speech signal based on the observation corrupted by additive noise. The female and male speech signals contaminated by white noise, pink noise and machine noise were measured respectively. More specifically, we created noisy air-conducted speech signals on a computer by mixing the original air-conducted speech signal measured in a noise-free environment with white and pink noises generated from a noise generator and machine noise recorded in advance. Furthermore, the bone-conducted speech signal was simultaneously measured by use of an acceleration sensor with the air-conducted speech signal.

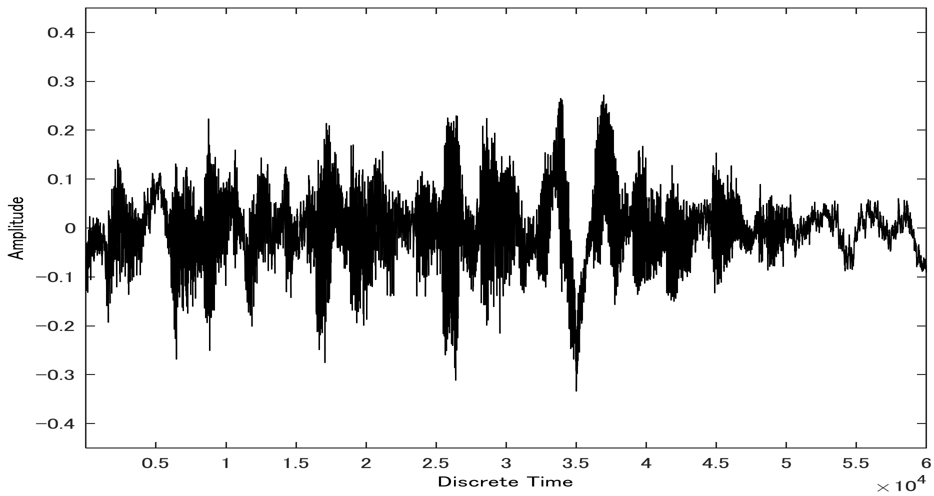

Figure 1: Noise-free male speech signal.

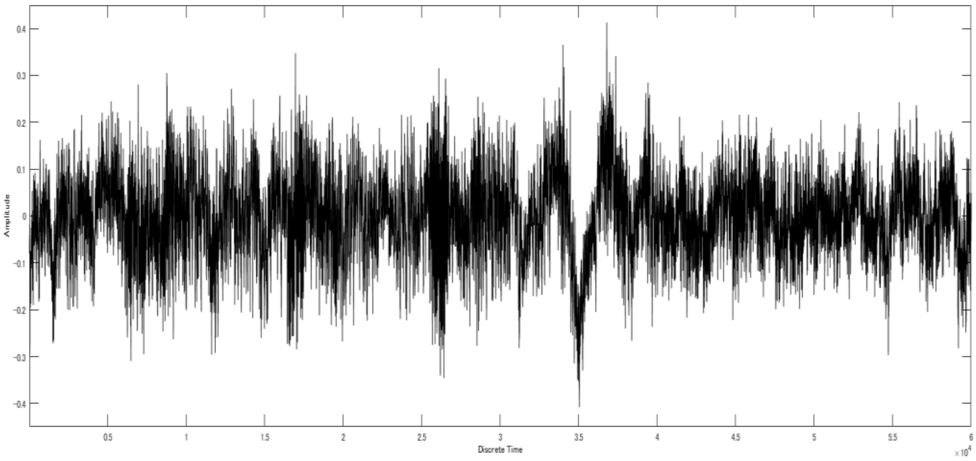

Figure 2: Noisy air-conducted speech signal mixed with machine noise.

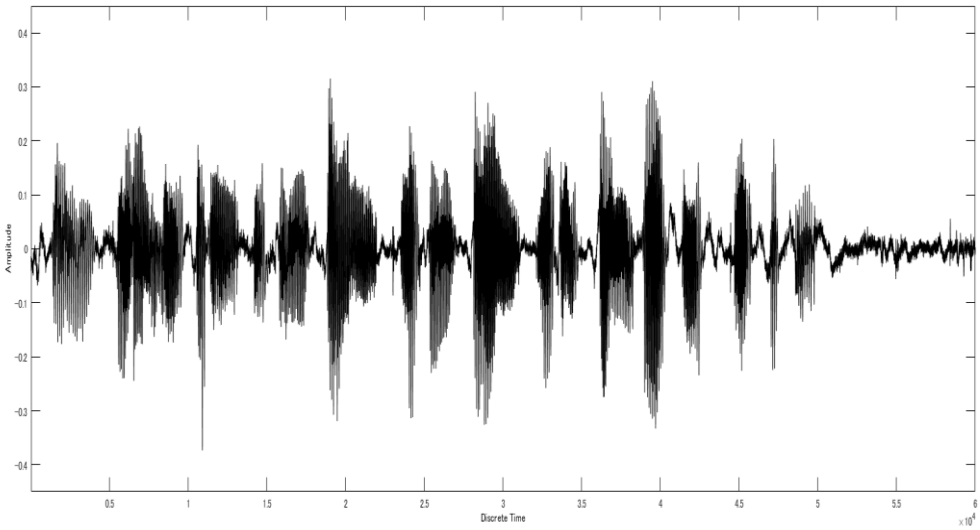

Figure 3: Observed wave of the bone-conducted speech signal.

As an example, the noise-free air-conducted male speech signal and the created noisy air-conducted observation signal using machine noise with the same amplitude as the noise-free speech signal are shown in Figures 1 and 2. Furthermore, the observed wave of the bone-conducted speech signal is shown in Figure 3.

The estimated result of the male speech signal by using the algorithm based on (9) is shown in Figure 4. For comparison, the estimated result of the male speech signal is shown in Figure 5. It is obvious that the effect of noise is not sufficiently removed by using the estimation algorithm based on only the observation of air-conducted speech signal, as compared with the proposed algorithm based on the observation of both air- and bone- conducted speech signals. By comparing Figures 4 and 5 with the original male speech signal shown in Figure 1, it is obvious that the proposed method can suppress the effects of the real machine noise better than the observation of only air-conducted speech signal.

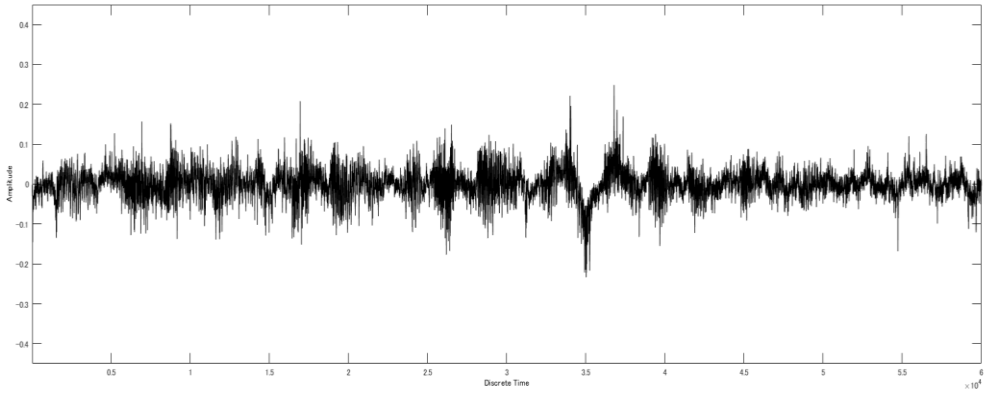

Figure 4: Estimated speech signal by use of the proposed method.

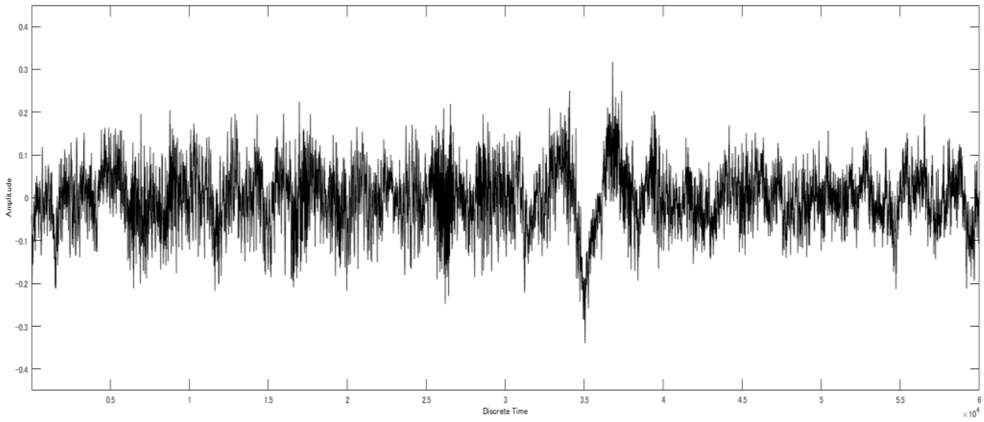

Figure 5: Estimated speech signal by use of the method based on only air-conducted observations.

4. CONCLUSIONS

In this paper, a new method to suppress noises in speech signals is proposed by jointly using the measured data of bone- and air- conducted speech signals. Furthermore, the effectiveness of the proposed algorithm has been confirmed by applying it to real speech signals and noises measured in an anechoic chamber of an acoustic building.

The proposed approach is quite different from the traditional standard techniques. However, we are still in an early stage of development, and a number of practical problems are yet to be investigated in the future. These include: (i) application to an even more diverse range of speech signals in actual noise environments, (ii) extension to cases with multi-noise sources, and (iii) finding an optimal number of expansion terms for the expansion-based probability expressions adopted.

The authors would like to thank Mr. Tomoya Shimono of the Prefectural University of Hiroshima and MEXT who in part supported this work by fund from the Grant-in-Aid for Scientific Research No.15K06116.

REFERENCES

1. Boll SF. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans Acoustics Speech and Signal Process*, **27**(2), 113-120, (1979).

2. Virag N. Single channel speech enhancement based on masking properties of the human auditory system, *IEEE Trans Speech Audio Process*, **7**(2), 126-137, (1999)

3. Kaneda Y, Ohga J. Adaptive microphone-array system for noise reduction. *IEEE Trans Acoustics Speech and Signal Process*. **34**(6), 1391-1400, (1986).

4. Kawamura A, Fujii K, Itoh Y, Fukui Y. A noise reduction method based on linear prediction analysis. *IEICE Trans Fundamentals*. **J85-A**(4), 415-423, (2002).

5. Kawamura A, Iiguni Y, Itoh Y. A noise reduction method based on linear prediction with variable step-size. *IEICE Trans Fundamentals*. **E88-A**(4), 855-861, (2005).

6. Gabrea M, Grivel E, Najim M. A single microphone Kalman filter-based noise canceller. *IEEE Signal Process Lett*. **6**(3), 55-57, (1999).

7. Kim W, Ko H. Noise variance estimation for Kalman filtering of noisy speech. IEICE Trans Inf & Syst. **E84-D**(1), 155-160, (2001).

8. Tanabe N, Furukawa T, Tsuji S. Robust noise suppression algorithm with the Kalman filter theory for white and colored disturbance. *IEICE Trans Fundamentals*. **E91-A**(3), 818-829, (2008).

9. Kalman RE. A new approach to linear filtering and prediction problems. *Trans. ASME, Series D, J. Basic Engineering*. **82**(1), 35-45, (1961).

10. Kalman RE, Buch RS. New results in linear filtering and prediction theory. *Trans ASME Series D, J Basic Engineering*. **83**(1), 95-108, (1961).

11. Ikuta A, Orimoto H, Adaptive Noise Suppression Algorithm for Speech Signal Based on Stochastic System Theory. *IEICE Trans on Fundamentals of Electronics, Communications and Computer Sciences*. **E94-A**(8), 1618-1627, (2011).

12. Julier, Simon J, Uhlmann, Jeffrey K. (1997). A new extension of the Kalman filter to nonlinear systems, *Int. Symp Aerospace/Defense Sensing, Simul. and Controls. Signal Processing, Sensor Fusion, and Target Recognition VI*, 182-193, (1997).

13. Ohta M, Yamada H. New methodological trials of dynamic state estimation for the noise and vibration environment system. *Acustica*. **55**(4): 199-212, (1984).

14. Ikuta A, Ohta M. A state estimation method of impulsive signal using digital filter under the existence of external noise and its application to room acoustics. *IEICE Trans Fundamentals*.; **E75-A**(8), 988-995, (1992).

15. Ohta M, Koizumi T. General statistical treatment of the response of a non-linear rectifying device to a stationary random input. *IEEE Trans Inf Theory*, **14**(4), 595-598, (1968).