PAPER Special Section on Robust Speech Processing in Realistic Environments

Multichannel Speech Enhancement Based on Generalized Gamma Prior Distribution with Its Online Adaptive Estimation

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

SUMMARY We present a multichannel speech enhancement method based on MAP speech spectral magnitude estimation using a generalized gamma model of speech prior distribution, where the model parameters are adapted from actual noisy speech in a frame-by-frame manner. The utilization of a more general prior distribution with its online adaptive estimation is shown to be effective for speech spectral estimation in noisy environments. Furthermore, the multi-channel information in terms of cross-channel statistics are shown to be useful to better adapt the prior distribution parameters to the actual observation, resulting in better performance of speech enhancement algorithm. We tested the proposed algorithm in an in-car speech database and obtained significant improvements of the speech recognition performance, particularly under non-stationary noise conditions such as music, air-conditioner and open window.

key words: *multi-channel speech enhancement, speech recognition, generalized gamma distribution, moment matching*

1. Introduction

In present days, speech recognition has been studied more and more in realistic environments such as home, office, hospital, car, bank, e.g., [1]–[4]. However, the robustness issue is still the main obstacle for the technique to be adopted in many real-life applications. To address this problem, many approaches have been developed in recent years. These methods can be separated into model-based [5]–[10] and data-driven [11]–[13].

The data-driven approaches, including the feature enhancement [11] and model compensation [12], [13], have shown to be very effective in specific environments, where the noise behavior can be learnt using large recorded databases but not quite convenient to apply in environments of high variety.

In contrast, the model-based approaches, using some knowledge to learn the short-term actual characteristics of the signal and the noise in order to reduce the latter from the given record, can be applied in more general cases.

Statistical speech enhancement [8] is a sub-direction of model-based methods and has shown good performances for the single channel approach. Previously, we have developed an approach based on speech estimation using adaptive generalized gamma distribution modeling on spectrum domain [9], [10] and were able to achieve better performance than conventional methods. The point of this method is that using more general prior distribution with online adaptation could more accurately estimate the signal and noise statistics and therefore improve the speech estimation, resulting in better performance of speech enhancement. Furthermore, the generalized gamma model is suitable in the implementation, yielding tractable-form solutions for the estimation and adaptation. In this work, we shall extend this method to the multi-channel case to improve the speech recognition performance.

It is well known that, the most conventional multi-channel speech enhancement method is the beamforming [14]. This method exploits spatial information such as direction of arrival of the sources in order to remove the interference signals. However, the performance of this method becomes severely degraded in many realistic cases including strong background noise, multiple sources, and "near-field" environments (such as in-car).

The multi-channel statistical speech enhancements were less studied in the literature. Two methods were proposed in [15], [16] and they obtained better performance than beamforming but those are limited for the incoherent noise fields (i.e. the uncorrelation of noise spectra in channels is assumed), which is not typical for most realistic noise environments. Moreover, the Gaussian model, that was shown to be non-optimal for speech modeling in single channel approaches, was adopted in [15], [16].

In this work, we will first develop a theoretical estimation framework for the multi-channel statistical speech enhancement for a general case of noise environment. Then we will use the spatial information in terms of cross-channel statistics to improve the parameter estimation of the generalized gamma model and consequently the speech enhancement compared to the single channel approach.

The organization of the rest of the paper is as follows. In Sect. 2 we will summarize (Reviewer 2.4) the multi-channel model of noisy speech in spectral domain. In Sect. 3 we will develop theoretical framework for speech spectral estimation based on multi-channel noisy speech signals. Section 4 will show how we can estimate the multi-channel statistics and use them for the system parameter estimation. Section 5 consists of an evaluation of the experiment followed by a overall summary of the work in Sect. 6.

Manuscript received July 9, 2007.
Manuscript revised September 14, 2007.

[†]The author is with Institute for Infocomm Research, 21 Heng-MuiKeng Terrace, Singapore 119613.
[††]The author is with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8601 Japan.
[†††]The author is with the Graduate School of Information Engineering, Meijo University, Nagoya-shi, 468–8502 Japan.
a) E-mail: hdtran@i2r.a-star.edu.sg

DOI: 10.1093/ietisy/e91–d.3.439

2. Multichannel Noisy Speech Model

We consider the additive model of multichannel signals in the STDFT domain

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

where $\mathbf{X} = [X_1, \ldots, X_D]^T$, $\mathbf{N} = [N_1, \ldots, N_D]^T$ are the vectors of the complex spectra of noisy and noise signals, $\mathbf{H} = [H_1, \ldots, H_D]^T$ is the vector of transfer functions, S is clean speech spectrum and $l = 1 : D$ is the microphone index. Couple (n,k) denotes the time-frequency index but will be omitted in Sects. 2 and 3.

The main points of our assumptions, what are different from previous works on the topic [15], [16], are as follows.

-Due to a possible change in the positions of speakers, we do not assume transfer functions to be constant in each frequency bin.

-The noise is assumed to be spatially coherent. This assumption widens the class of noise not only limited by diffused background noise and is one important point of the proposed method.

The noise spectral components is assumed to follow the zero-mean multi-variable Gaussian distribution with a full covariance matrix

$$\begin{aligned} p(\mathbf{N}_R) &= \frac{1}{2\pi^{D/2}\det(2\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{N}_R^T \mathbf{C_n}^{-1} \mathbf{N}_R\right\}, \\ p(\mathbf{N}_I) &= \frac{1}{2\pi^{D/2}\det(2\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{N}_I^T \mathbf{C_n}^{-1} \mathbf{N}_I\right\}, \end{aligned} \tag{2}$$

the noise spectral components are assumed to be statistically independent, i.e.

$$p(\mathbf{N}_R, \mathbf{N}_I) = p(\mathbf{N}_R)\, p(\mathbf{N}_I), \tag{3}$$

where $(.)_R$, and $(.)_I$ denote the real and imaginary parts of the complex spectrum, respectively and $\mathbf{C_n}$ is half of the covariance matrix (real and symmetrical).

-Speech spectral magnitude $|S|$ follows a generalized gamma distribution given as

$$p(|S|) = \frac{b^a}{\Gamma(a)\,\sigma_S} \left(\frac{|S|}{\sigma_S}\right)^{La-1} \exp\left[-b\left(\frac{|S|}{\sigma_S}\right)^L\right], \tag{4}$$

where σ_S^2 denotes the variance of speech spectrum. (a, b, L) are distribution parameters but the system has two remaining free parameters due to the normalization $\left\langle |S|^2 \right\rangle = \sigma_S^2$, where $\langle . \rangle$ denotes the expectation.

The reasons why the generalized gamma distribution was employed are two. Firstly, this distribution is superset of many distribution models, what were used for speech modeling such as Gaussian model of spectral components, generalized Gaussian model, gamma distribution of speech magnitude and therefore is expected to better model the speech prior distribution [9]. Secondly, the model is suitable to be adapted from actual noisy speech as this was shown in [9], [10]. As the prior distribution is better estimated using this model, the speech estimation and ASR performance are improved. We have added the description and explanation as was suggested. The generalized distribution was used in our previous studies for single-channel approaches [9], [10] where the model could achieve superior peformances compared to the conventional ones. In this work, we will show how the multi-channel statistical information can be used for this model to further improve the performance of speech recognition systems.

3. Speech Spectral Magnitude Estimation Based on Generalized Gamma Model Using Multi-Channel Microphone

In this section we develop the speech spectral magnitude estimation from multi-channel noisy speeches under a general noise environment.

Similar to the single channel approaches, given observations in terms of multi-channel noisy speech signals $\mathbf{X}$, the statistical speech enhancement estimate the speech spectrum; particularly the speech spectral magnitude using a Bayesian estimator such as Minimum Mean Square Error (MMSE) or Maximum a Posterior Probability (MAP).

For the proposed generalized gamma model, the MAP estimation method is found to be more preferable due to the effectiveness as well as the simplicity in the implementation [9], [10]. The multichannel MAP estimation equation

$$\left|\hat{S}\right| = \arg\max_{|S|} \left[p(|S|\,|\mathbf{X})\right] \tag{5}$$

can be denoted via the Bayesian formula as

$$\frac{\partial}{\partial |S|} \left\{\log\left[p(\mathbf{X}|\,|S|)\right] + \log\left[p(|S|)\right]\right\} = 0. \tag{6}$$

As the prior PDF of speech magnitude is determined by (4), the conditional PDF component in (6) needs to be derived from (1)–(3). In contrast to the methods proposed in [15], [16], we derive the estimation for the general case of a spatially coherent noise using generalized gamma model of speech prior distribution.

3.1 Derivation of Multichannel Rician Distribution

From (3), $p(\mathbf{X}|\,|S|)$ can be factorized as

$$p(\mathbf{X}|S) = p(\mathbf{X}_R|S_R, S_I)\, p(\mathbf{X}_I|S_R, S_I). \tag{7}$$

Note that here, we consider the transfer functions as deterministic variables, which will be estimated from observations. Two terms in right side of (6) can be denoted using (1)–(3)

$$\begin{aligned} &p(\mathbf{X}_R|S_R, S_I) \\ &= \frac{1}{2\pi^{D/2}\det(\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{Y}_R^T \mathbf{C_n^{-1}} \mathbf{Y}_R\right\}, \\ &p(\mathbf{X}_I|S_R, S_I) \\ &= \frac{1}{2\pi^{D/2}\det(\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{Y}_I^T \mathbf{C_n^{-1}} \mathbf{Y}_I\right\}, \end{aligned} \tag{8}$$

where

$$\begin{aligned}\mathbf{Y}_R &= \mathbf{X}_R - \mathbf{H}_R S_R + \mathbf{H}_I S_I,\\ \mathbf{Y}_I &= \mathbf{X}_I - \mathbf{H}_R S_I - \mathbf{H}_I S_R.\end{aligned} \tag{9}$$

Substituting (8) and (9) into (7) yields the conditional distribution of the complex spectrum of noisy speech expressed as

$$\begin{aligned}&p(\mathbf{X}|S)\\ &= Q(\mathbf{X}) \exp\left\{-\left[\begin{array}{l}\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}\left(S_R^2+S_I^2\right)-\\ -2Re\left(\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}\right)S_R-\\ -2Im\left(\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}\right)S_I\end{array}\right]\right\},\end{aligned} \tag{10}$$

where $Q(\mathbf{X})$ is independent of S and this term will be reduced with further estimation. Since $\mathbf{C_n}$ is real and symmetrical the term $\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}$ is real. The conditional distribution (10) can be transformed into magnitude and phase, using Jacobian transform, yielding

$$\begin{aligned}&p(\mathbf{X}|\varphi_S, |S|)\\ &\triangleq \exp\left\{\begin{array}{l}-\left(\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}\right)|\mathbf{S}|^2-\\ -2|\mathbf{S}|\left|\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}\right|^2\cos(\varphi_\mathbf{X}-\varphi_S)\end{array}\right\},\end{aligned} \tag{11}$$

where φ_S denotes the phase of the clean speech spectrum and φ_X- the phase of $\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}$. Integrating (11) over φ_S, we obtain the conditional distribution of noisy speech magnitude as a multichannel version of the Rician distribution [9].

$$\begin{aligned}&p(\mathbf{X}||S|)\\ &\triangleq \exp\left(-\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}|\mathbf{S}|^2\right)I_0\left(2|\mathbf{S}|\left|\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}\right|\right).\end{aligned} \tag{12}$$

Here, I_0 is the Bessel function of the first kind, which can be approximated by

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}}e^x, x > 0. \tag{13}$$

The first term in (6) is then derived as

$$\frac{\partial}{\partial|S|}\{\log[p(\mathbf{X}||S|)]\} = -\frac{2|S|}{U_n^2} - \frac{1}{2S} + \frac{\frac{|\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}|}{\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}}}{U_n^2}, \tag{14}$$

where

$$U_n^2 = \frac{1}{\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}}. \tag{15}$$

3.2 Estimation Equation

Now we are getting the estimation equation. From (4), the second term in (6) is expressed by

$$\frac{\partial}{\partial|S|}\{\log[p(|S|)]\} = \frac{(La-1)}{|S|} - Lb\frac{|S|}{\sigma_S^{L-1}}. \tag{16}$$

Substituting (14) and (16) into (6) yields the estimation equation,

$$\frac{1}{|S|}(La-1.5) - |S|\left(\frac{Lb}{\sigma_S^{L-1}} + \frac{2}{U_n^2}\right) + \frac{\frac{|\mathbf{H}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}|}{\mathbf{H}^T\mathbf{C}_\mathbf{n}^{-1}H}}{U_n^2} = 0, \tag{17}$$

which generally can be solved by the Newton-Raphson method [9].

A special case, what we experimentally found to be very effective is when $L = 2$ [10] yielding a closed-form solution and therefore is suitable in the implementation. The estimation equation is then derived as

$$-G^2 + \frac{G}{\left(1+\frac{b}{\xi}\right)} + \frac{4a-3}{G} = 0. \tag{18}$$

Here the gain function is

$$G = \frac{\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{H}}{\left|\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}\right|}|S|, \tag{19}$$

and the generalized a priori and posteriori SNR (frame SNRs) are defined as

$$\xi = \frac{\sigma_S^2}{U_n^2}, \gamma = \left|\bar{\mathbf{H}}^T\mathbf{C}_\mathbf{n}^{-1}\mathbf{X}\right|^2. \tag{20}$$

Now the speech spectral magnitude can be estimated using (17) or (18). The remaining problem is to estimate the system parameters, i.e. the noise covariance, the transfer function, the frame SNR and the prior distribution parameters.

4. Online Adaptive Parameter Estimation

Similar to our single channel approach, the prior distribution parameter is estimated in actual (i.e. from noisy speech) and updated in an online (i.e. frame by frame) manner. However, the main point of this paper is that the cross-channel statistics, are used to adapt the prior distribution parameter closest to the actual observations. Using multi-channel statistical information, we will be able to faster react in the change of the ambient noise and therefore more accurately estimate the system parameter and consequently improve the performance of speech enhancement, which is the key idea behind our proposed approach.

The block diagram of system parameters estimation is shown in Fig. 1. The noise covariance, speech power

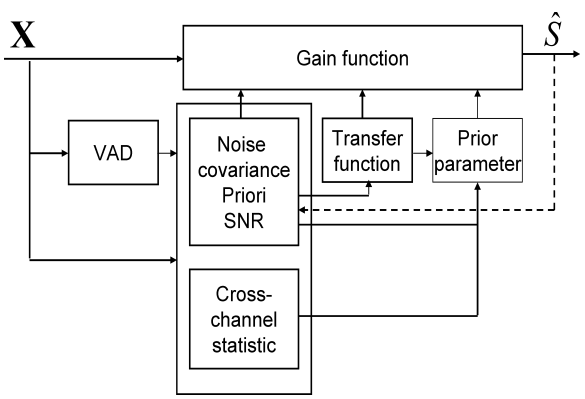

Fig. 1 Block-diagram processing.

and frame SNR (i.e. a priori and posteriori SNRs) are estimated from input noisy speech. These statistic measurements are updated using a multi-channel Voice Activity Detection which distinguishes the target speech and noise components. The transfer function is then estimated and updated using the frame SNRs. The frame SNRs are also smoothed by taking a feedback from estimated clean speech.

4.1 Noise Covariance, Frame SNRs and Transfer Functions Estimation

The noise covariance is initially estimated in each frequency bin using the first 250 ms of observations. Then it is recursively updated, using voice activity detection (VAD),

$$\mathbf{C_n}(n,k) = \begin{cases} \alpha \mathbf{C_n}(n-1,k) + (1-\alpha) Re\left[\bar{\mathbf{X}}^T \mathbf{X}(n,k)\right] & \text{D1}, \\ \mathbf{C_n}(n-1,k) & \text{D0} \end{cases} \tag{21}$$

where $D1$ and $D0$ are hypotheses of speech presence and absence, α is a smoothing coefficient. Differently from those methods proposed in [15], [16], we estimate the transfer function via priori SNR (i.e., short-term statistics). Not losing the generality, we assume

$$\mathbf{H}_1 = 1. \tag{22}$$

The magnitude of the transfer function at microphone i is determined and smoothed as

$$\widehat{|\mathbf{H}_i(n,k)|} = \begin{cases} \chi |\mathbf{H}_i(n-1,k)| + (1-\chi) \sqrt{\frac{\xi_i(n,k)}{\xi_1(n,k)} \frac{\sigma^2_{n_1}(n,k)}{\sigma^2_{n_i}(n,k)}} & \text{D1} \\ |\mathbf{H}_i(n-1,k)|, & \text{D0} \end{cases} \tag{23}$$

where χ is a smoothing coefficient.

The priori SNR in the i-channel $\xi_i = \frac{|H_i|^2 \sigma_S^2}{\sigma^2_{n_i}}$ is estimated by the decision-directed method [5].

The phase of the transfer function is estimated using the covariance matrix relationship

$$\mathbf{C}_x(n,k) = \bar{\mathbf{H}}^T(n,k)\, \mathbf{H}(n,k)\, \sigma_S(n,k)^2 + \mathbf{C}_n(n,k)(1+j), \tag{24}$$

where $\mathbf{C}_x$ is the noisy speech covariance matrix, which is initially estimated as same as $\mathbf{C}_n$ and later is updated as in (21) but without using VAD.

$$\mathbf{C}_x(n,k) = \alpha \mathbf{C}_x(n-1,k) + (1-\alpha)\, \bar{\mathbf{X}}^T(n,k)\, \mathbf{X}(n,k) \tag{25}$$

Taking into account (24), the phase of the transfer function can be estimated using the first column of $\mathbf{C}_x$.

4.2 Prior Distribution Parameter Estimation

Now we will discuss the most important part of the system estimation: the prior distribution parameters estimation. Similar to the single channel case, we use the moment matching method to estimate and adapt the prior distribution from actual noisy speech. However, the key point here is that the multi-channel statistics are used in addition in order to better and faster "match" the prior distribution.

Denote the noisy speech power as

$$|X_i|^2 = |H_i|^2 |S|^2 + |N_i|^2 + 2|H_i||S||N_i| \cos(\Delta\phi_i), \tag{26}$$

where $\Delta\phi_i$ is the phase difference, which is assumed to follow a uniform distribution [9].

Taking into account the independence of the phase and the magnitude of the noise spectrum, the cross-channel correlations of noisy speech power are expressed as

$$\begin{aligned} \left\langle |X_1|^2 |X_i|^2 \right\rangle &= |H_i|^2 \left\langle |S|^4 \right\rangle + \left\langle |N_1|^2 |N_i|^2 \right\rangle \\ &+ \left\langle |S|^2 \right\rangle \left(|H_i|^2 \left\langle |N_1|^2 \right\rangle + \left\langle |N_i|^2 \right\rangle + 4|H_i| |\langle N_1 N_i \rangle| \right), \end{aligned} \tag{27}$$

where $\langle . \rangle$ denotes the expectation operator.

At the same time, the product of speech powers can be expressed by

$$\begin{aligned} \left\langle |X_1|^2 \right\rangle \left\langle |X_i|^2 \right\rangle &= |H_i|^2 \left(\left\langle |S|^2 \right\rangle \right)^2 \\ &+ \left\langle |N_1|^2 \right\rangle \left\langle |N_i|^2 \right\rangle \\ &+ \left\langle |S|^2 \right\rangle \left(|H_i|^2 \left\langle |N_1|^2 \right\rangle + \left\langle |N_i|^2 \right\rangle \right). \end{aligned} \tag{28}$$

Since all components in the lower term of (27) and (28) can be determined from estimated transfer function, noise covariance matrix and priori SNR estimations, the ratio of the fourth to the square of the second-moments of speech magnitude $|S|$ can be determined in an analytical form (at each time-frequency index). In general case, this "matching" ratio is given in following form

$$\frac{\left\langle |S|^4 \right\rangle}{\left\langle |S|^2 \right\rangle^2} = \frac{\Gamma\left(a + \frac{4}{L}\right) \Gamma(a)}{\Gamma\left(a + \frac{2}{L}\right)^2}, \tag{29}$$

what can be simplified after taking logarithm

$$\begin{aligned} \log\left(\frac{\left\langle |S|^4 \right\rangle}{\left\langle |S|^2 \right\rangle^2} \right) &= \log\left(\Gamma\left(a + \tfrac{4}{L}\right) \right) \\ &+ \log(\Gamma(a)) - 2\log\left(\Gamma\left(a + \tfrac{2}{L}\right) \right). \end{aligned} \tag{30}$$

In general, both the prior distribution parameters can be estimated using the moment matching method. However, in order to minimize the trade-off between performance and computational cost, we do not optimize the order parameter L but select this parameter experimentally. In this case, this leads to much more tractable-form solution for estimating and adapting one parameter a.

Thanks to Gergo Nemes for his approximation of gamma function [17],

$$\log(\Gamma(z)) \approx \tfrac{1}{2}(\log(2\pi) - \log(z)) + z\left(\log\left(z + \frac{1}{12z - \frac{1}{10z}}\right) - 1\right), \quad (31)$$

the matching equation in general case can be denoted as a compact form to be solved.

For the special case $L = 2$, what we found to be very effective through the experiments, this yields a simple analytical form

$$\left\langle |S|^4 \right\rangle = \frac{a(a+1)}{b^2}\left(\left\langle |S|^2 \right\rangle\right)^2, \quad (32)$$

yielding an easy solution, suitable for the implementation,

$$a = \frac{1}{\frac{\langle S^4 \rangle}{(\langle S^2 \rangle)^2} - 1}. \quad (33)$$

Note that here we take into account the normalization $\left\langle |S|^2 \right\rangle = \sigma_S^2$ which implies the relationship $a = b$.

Now what remains is to estimate the cross-channel correlations of noisy speech powers and noise powers. As both of these statistics are initially estimated using first 250 ms, the noisy speech component is updated by recursive moving averages without using VAD

$$\begin{aligned} &\left\langle |X_1(n,k)|^2 |X_i(n,k)|^2 \right\rangle \\ &= \alpha \left\langle |X_1(n-1,k)|^2 |X_i(n-1,k)|^2 \right\rangle \\ &+ (1-\alpha)|X_1(n,k)|^2 |X_i(n,k)|^2. \end{aligned} \quad (34)$$

Alternatively, the noise statistic is updated only in the non-speech segment.

Given the cross-channel statistics the prior distribution parameter can be determined using (29)–(31). Finally, we smooth the estimated parameter to avoid any spike in the estimation

$$a(n,k) = \mu a(n-1,k) + (1-\mu)\, a(n,k), \quad (35)$$

where μ is a smoothing coefficient.

4.3 Voice Activity Detection

For VAD, we assume that the nearest microphone (i.e. less noisy) is known in advance and will be indexed by number-1. In addition to the conventional power feature given from this channel, we use the cross-channel correlation coefficient ρ calculated from each frame to determine VAD.

$$\rho(n) = \frac{1}{D-1}\sum_{i=2}^{D} \sqrt{\frac{|\langle X_1(n) X_i(n)\rangle|}{\left\langle |X_1^2(n)| \right\rangle \left\langle |X_i^2(n)| \right\rangle}} \quad (36)$$

Here we assume that the channels have a higher correlation in the target signal durations.

We note that the cross-channel correlation in (36) is similar to the coherent function used in a previous VAD proposed in [18]. But here we combined this measurement with the frame power distance to compensate some errors caused by possible coherent noise segments.

The resulting VAD is expressed as

$$VAD(n) = \begin{cases} 1 & if\ \ |\Delta(n)| > \gamma_1\ \&\ \rho(\mathrm{n}) > \gamma_2 \\ 0 & otherwise \end{cases}, \quad (37)$$

where γ_1 and γ_2 are constant boosting factors. Currently $\gamma_1 = 3$ and $\gamma_2 = 0.3$ are used. The energy distance Δ is the ratio of current frame energy $E_1(n)$ to the stored-in-memory noise energy and is calculated in decibels. Note that, the VAD in our system is used just to update the noise statistics but not meant to drop the noise frames from speech recognition system. Although the comparison of VAD algorithms are out of the scope of this paper we found that our method can quite well distinguish the target signal from background noise even in very low SNR conditions. The VAD accuracy is critically important for the low-SNR and non-stationary noise conditions as it could help the recognizer to reduce the highly possible recognition errors caused by noise frames.

5. Experiment

We evaluate the proposed algorithm in CIAIR in-car speech corpus [19]. The data collection setup is shown in Fig. 2. The training data for HMMs consists of 293 speakers with 7,000 phonetically balanced sentences. The test data is a set of 50 isolated words recorded from other 6 speakers in 15 different driving conditions.

In the ASR implementation, 16-kHz sampling signals from two microphones, attached to the ceiling positions (i.e. mic.5-6 in Fig. 2), were used. For the reference single channel speech enhancement methods, the nearest-to-driver ceiling microphone (i.e. mic.6 in Fig. 2) was used. The Hamming window of length of 20 ms with 50% overlap was applied in FFT. The 25-feature configuration of 12MFCC+12delta-MFCC+log-energy was adopted in the implementation.

Figure 1 shows the block diagram of processing. Given the multichannel noisy signals, VAD is performed using (37). The noise covariance, priori SNR and cross-channel statistics are estimated using recursive averages. The transfer function is estimated using (23). The prior parameters are estimated and updated using (29) or (33) and (35). Then the speech spectral magnitude is estimated by (17) or (18).

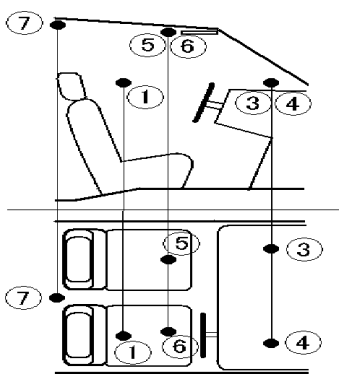

Fig. 2 Experiment setup.

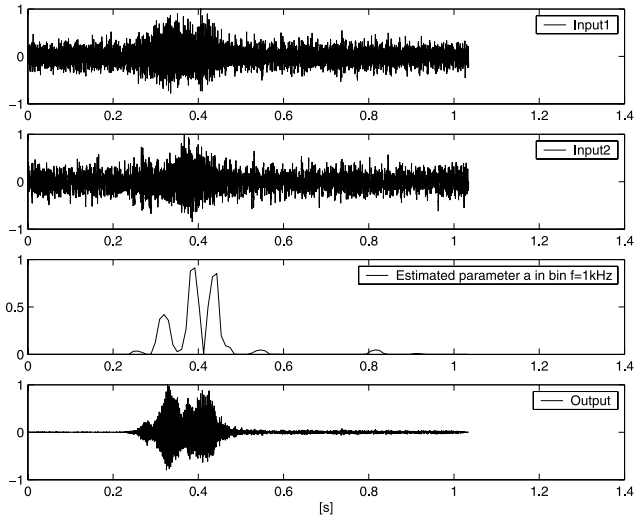

Fig. 3 Input 2-channel noisy signals (top two plots), example of estimated prior distribution parameter in bin f = 1 kHz (third plot) and output waveform (bottom).

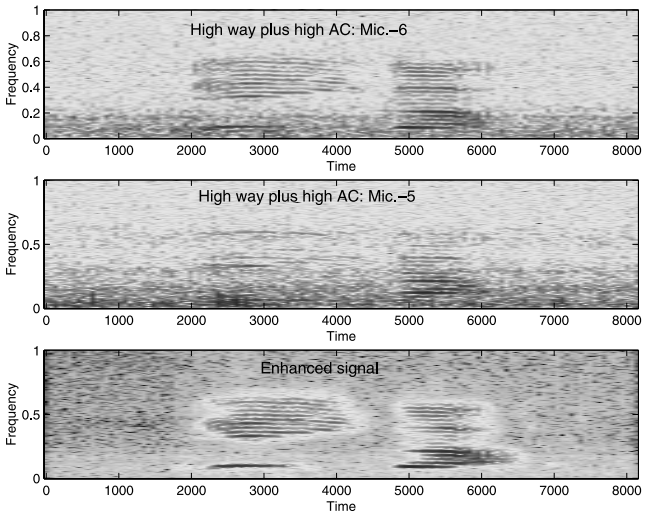

Fig. 4 Examples of noisy and enhanced spectrogram for "High air-con in highway". From top to bottom: noisy spectrogram in mic.6, noisy spectrogram in mic.5 and spectrogram of the enhanced signal. The frequency scale is 1 : 8000 [Hz], the time scale is 1 : 16000 [s].

Finally, the phase adding and overlap and add are used to re-synthesize the enhanced sounds. The smoothing coefficients are chosen by hearing the output sounds. Currently, $\alpha = \beta = \chi = 0.9$, and $\mu = 0.6$ are used and the performances seem to not very sensitive to the smoothing parameters.

We note that, in general, speech recognition can be performed without re-synthesizing the enhanced signals as the MFCC can be calculated directly from speech spectral estimations and the same results should be obtained. However, the sound re-synthesis could help us to analyze the processing by hearing in order to tune some smoothing coefficients more easily. Moreover, we can assume that the enhanced sound output is additionally required for possible human-human interaction and therefore the re-synthesis was implemented.

An example of waveform processing for a noisy speech signals recorded in "CD playing in high way" driving condition is plotted in Fig. 3. The noisy signal are in two upper plots, the third plot show the estimated prior parameter and the last plot is the enhanced signal in a waveform. We can see that the optimal prior distribution parameter is varying in the speech segments. The multi-channel VAD can well distinguish the target speech to the background noise and the output signal is much better than the original ones.

Examples of processing in spectral domain for the most severe non-stationary noise conditions, i.e. the "CD playing in highway", the "High air-con level in highway" and the "Window opening in highway" are illustrated in Figs. 4–6.

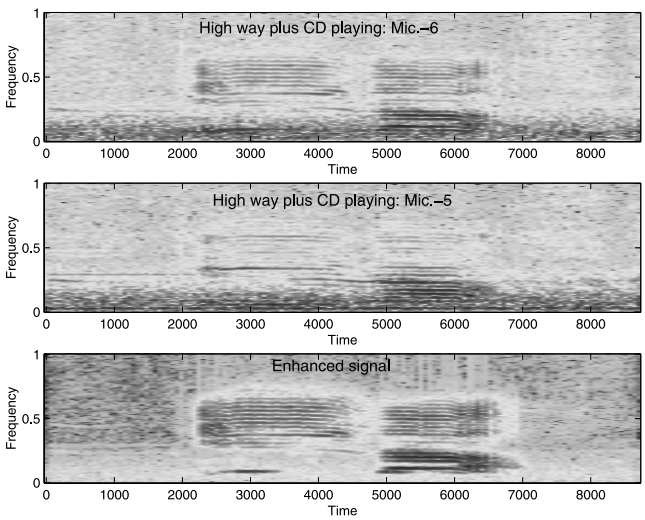

Fig. 5 Examples of noisy and enhanced spectogram for "CD playing in highway".

5.1 Speech Recognition Evaluation Comparison

For the speech recognition evaluation, we implemented the following methods:

1. The nearest-to-driver ceiling microphone (i.e. mic.6 in Fig. 2) without enhancement;
2. Ephraim-Malah Log-Spectral Magnitude estimation method based on Gaussian model (LSA) [6];
3. Our previous single-channel version using generalized gamma modeling (GG) [9];
4. The most recent multi-channel speech enhancement based on psychoacoustic motivation [16];
5. The proposed multi-channel method with closed form solution (i.e. $L = 2$) using two channels in ceiling positions.

The overall results of speech recognition are shown in Fig. 7. All the speech enhancement method could greatly improve the accuracy of speech recognition more than 10% in absolute accuracy rate compared to the performance in the nearest microphone. The proposed multichannel method

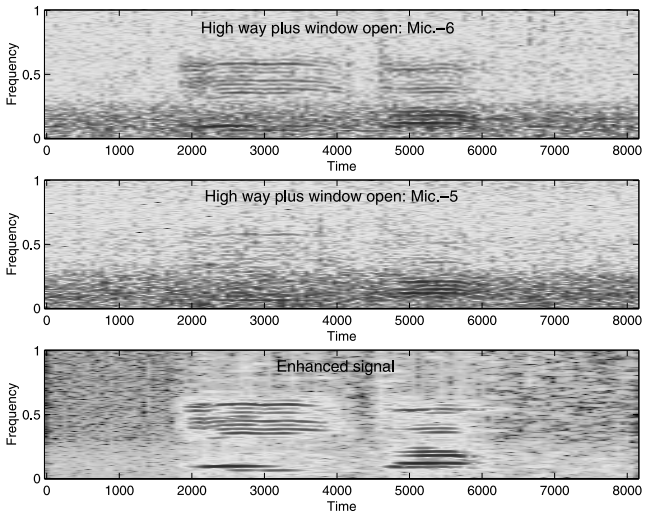

Fig. 6 Examples of noisy and enhanced spectrogram for "Window open in highway".

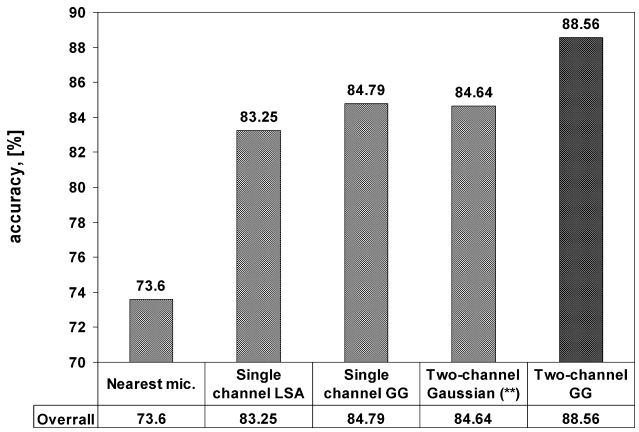

Fig. 7 Speech recognition overall performances evaluated on CIAIR in-car database, Two-channel Gaussian (**) is the multi-channel speech enhancement based on psychoacoustic motivation [16].

shown to have a performance improvement over other methods by approximately 16%. The superiority of generalized gamma modeling to the conventional Gaussian model (LSA method) is confirmed again. The reason is clear: using the more general distribution with its online adaptation improves the accuracy of prior distribution modeling which is realized in the MAP estimation. The single-channel method using generalized gamma modeling achieved almost the same result as the two-channel method using Gaussian model and psychoacoustic knowledge [16]. However, the use of additional information in terms of multi-channel statistics seems to better adapt the model distribution and yield less distortion and therefore outperform the single channel with approximately 4% improvement.

5.2 Significant Improvements under Severely Non-stationary Noise Conditions

The proposed method is especially effective under non-stationary noise conditions. Table 1 shows the results for the cases of driving along an express way with a CD playing, high air-conditioner (AC) and open window. The improvement of the proposed method for these conditions are approximately 10%, 40% and 44% compared to the nearest microphone. The proposed method greatly outperformed other methods in the high AC and window opening cases. This can be explained as follows. The multi-channel method could react faster on the change of signal and noise statistics, moreover this also better distinguishes the target speech and other noise sources which result in better performance of speech recognition. The online parameter estimation can also be considered as an optimization of the gain function, which controls the trade-off between noise reduction and distortion, resulting in the best speech recognition performances.

Table 1 Speech recognition rate on expressway under several driving conditions [%].

Method	Nearest mic	1-ch LSA	1-ch GG	2-ch PA	2-ch GG
CD playing	82.27	82.61	85.62	83.96	92.98
High AC	51.00	87.33	89.00	88.67	91.67
Window open	42.67	76.67	78.33	77.33	86.33

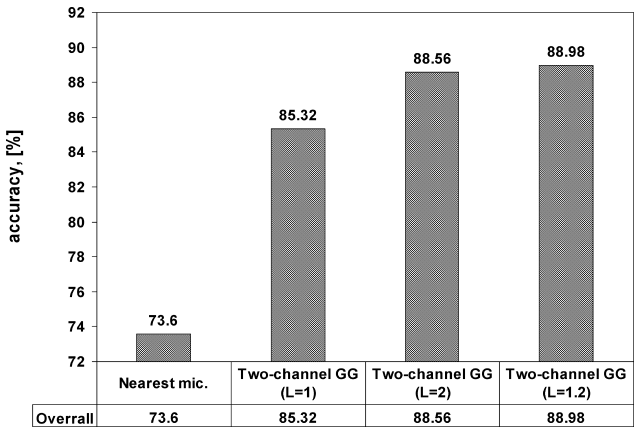

Fig. 8 Speech recognition overall performances: closed-form ($L = 2$) vs. iterative solutions.

5.3 Closed-Form vs. Iterative Solution

In this section we compare the proposed method on the selection of power order L, which lead to close-form solution (more easy to implement) for the case $L = 2$ or iterative solution for other cases. For the case $L = 1$ the Gain function is given analytically but the parameter estimation is given only by the iterative solution. As mentioned in previous sections, we will not discuss the automatic optimization of order L due to a to much more higher computation cost. Figure 8 shows the performance comparison for three cases of L between 1 and 2: $L = 1$, $L = 2$ and $L = 1.2$. This interval is experimentally found to be most effective and robust for the approach. We can see that, the "close-form solution" ($L = 2$) is better than iterative $L = 1$ but little worse than

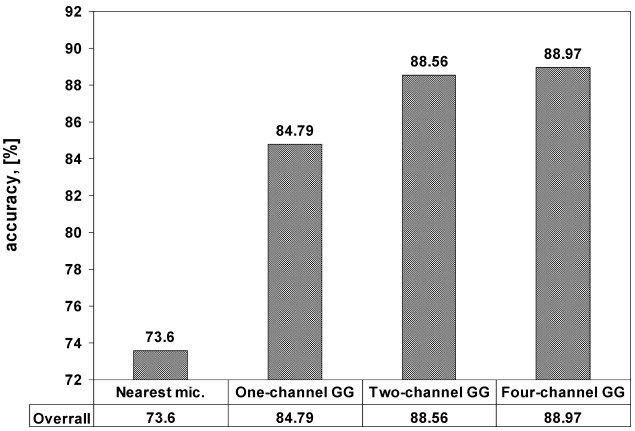

Fig. 9 Speech recognition overall performances: two-channel vs. four-channel.

$L = 1.2$. However, the last solution needs two iterative processes for estimating the gain function in (17) and tuning the prior parameter in (29) and therefore is much "heavier" than the "close-form solution" ($L = 2$). Therefore we recommend using this parameter in a general case.

5.4 2-Channel vs. 4-Channel

An arising question would be how would the performance be improved by using more microphone signals? To do this comparison for the in-car corpus we implement the proposed method one more time using four channels, two ceiling (i.e. mic.5 and mic.6) and two in-front (i.e. mic.3 and mic.4). Figure 9 compares the proposed method (with "close-form solution") for the three cases: single channel, two-channel and four-channel. We can see that the performance improvement has improved quite significantly when switching from single channel to the two-channel but there was no significant improvement when switching from two-channel to the four-channel. It seems that the additional spatial information is important but should be taken from right position. The choice of the number of microphones to be used should be specified for each environment. For the in-car case, the number of directional noise sources is quite few. The engine, air-con and highway noises are quite diffused and stationary although they might be in a high power level. Only the music-CD and window opening conditions contain directional noise sources. Another factor is the trade-off between the performance improvement and the computational cost. For the in-car, the two-channel approach is recommended.

6. Conclusions

This study has demonstrated the effectiveness of using more general prior distribution with online adaptation for the multichannel speech enhancement. The accuracy of prior distribution modeling using multichannel observations is the key point, which is realized in MAP speech spectral magnitude estimation. The experimental results show the superiority of the proposed method under non-stationary noise conditions.

Acknowledgement

This work has been partially supported by the MEXT leading project.

References

[1] A. Betkowska, K. Shinoda, and S. Furui, "Robust speech recognition using factorial HMMs for home environments," EURASIP Journal on Advances in Signal Processing, vol.2007, Article ID 20593, 9 pages, 2007. doi:10.1155/2007/20593
[2] M.A. Grasso, "The long-term adoption of speech recognition in medical applications," Proc. 16th IEEE Symposium on Computer-Based Medical Systems (CBMS 2003), pp.257–262, 2003.
[3] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," Proc. INTERSPEECH, ICSLP, pp.1213–1216, Pittsburgh, PA, USA, 2006.
[4] A. Acero, Acoustical and environmental robustness in automatic speech recognition, Kluwer Academic Publishers, 1993.
[5] 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.
[6] Y. Ephraim and D. Malah, "Speech enhancement using MMSE log-spectral amplitude estimations," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-33, no.2, pp.443–445, 1985.
[7] P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim suppression rule for speech enhancement," IEEE Workshop on Statistical Signal Processing, 2001.
[8] R. Martin, "Statistical methods for enhancement of noisy speech," Proc. IWAENC, Kyoto, 2003.
[9] T.H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," Proc. ICASSP, Philadelphia, USA, 2005.
[10] T.H. Dat, K. Takeda, and F. Itakura, "Gamma modeling of speech power and its on-line estimation for statistical speech enhancement," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.1040–1049, March 2006.
[11] W. Li, T. Shinde, H. Fujimura, C. Miyajima, T. Nishino, K. Itou, K. Takeda, and F. Itakura, "Multiple regression of log spectra for in-car speech recognition using multiple distributed microphones," IEICE Trans. Inf. & Syst., vol.E88-D, no.3 pp.384–390, March 2005.
[12] X. Cui and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," IEEE Trans. Acoust. Speech Signal Process., vol.13, no.6, pp.1161–1171, 2005.
[13] H. Shen, Q. Li, J. Guo, and G. Liu, "Model-based feature compensation for robust speech recognition," Fundam. Inf., vol.72, no.4, pp.529–539, Dec. 2006.
[14] D. Ward and M. Brandstein, Microphone Arrays: Signal Processing Techniques and Applications, Springer, ISBN 3540419535, 2001.
[15] T. Lotter, C. Benien, and P. Vary, "Multichannel direction-independent speech enhancement using spectral amplitude estimation," EURASIP Journal on Applied Signal Processing, vol.11, pp.1147–1156, 2003.
[16] J. Rosca, R. Balan, and C. Beaugeant, "Multi-channel psychoacoustically motivated speech enhancement," Proc. ICASSP, Hong Kong, 2003.
[17] V.T. Toth, "Programmable calculators: Calculators and the gamma function," http://www.rskey.org/gamma.htm
[18] R. Le Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," Speech Commun., vol.16, pp.245–254, 1995.

[19] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and F. Itakura, "Construction and evaluation of a large in-car speech corpus," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.553–561, March 2005.

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 2000, he received the PhD degree of Physic-Mathematical Science 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 fellow at the Itakura and Takeda Labs, Nagoya University. Since 2005 he is a Senior Research Fellow at Institute for Infocomm Research, Singapore. His research interest is including acoustic and speech signal processing, neural signal processing and machine learning. He served as a reviewer of international journals and conferences, including IEEE Transaction on Audio, Speech and Language Processing and Neurocomputing.

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 Information Science at Nagoya University.

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. He worked on isolated word recognition in the Acoustics Research Department of Bell Labs under James Flanagan from 1973 to 1975. 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.