A Priori SNR Estimation Using Air- and Bone-Conduction Microphones

Ho Seon Shin, Tim Fingscheidt, *Senior Member, IEEE*, and Hong-Goo Kang, *Member, IEEE*

***Abstract*—This paper proposes an *a priori* signal-to-noise ratio (SNR) estimator using an air-conduction (AC) and a bone-conduction (BC) microphone. Among various ways of combining AC and BC microphones for speech enhancement, it is shown that the total enhancement performance can be maximized if the BC microphone is utilized for estimating the power spectral density (PSD) of the desired speech signal. Considering the fact that a small deviation in the speech PSD estimation process brings severe spectral distortion, this paper focuses on controlling weighting factors while estimating the *a priori* SNR with the decision-directed approach framework. The time–frequency varying weighting factor that is determined by taking a minimum mean square error criterion improves the capability of eliminating residual noise and minimizing speech distortion. Since the weighting factors are also adjusted by measuring the usefulness of the AC and BC microphones, the proposed approach is suitable for tracking the parameter even if the characteristic of environment changes rapidly. The simulation results confirm the superiority of the proposed algorithm to conventional algorithms in high noise environments.**

***Index Terms*— *A priori* signal-to-noise ratio (SNR) , bone-conduction (BC) microphone, speech enhancement.**

I. Introduction

THE increase of mobile speech communication in noisy environments asks for advanced speech enhancement technology. Recently, speech enhancement for public safety speech communication systems that are operated in extremely high noise environments is being a new case, because keeping intelligibility or maintaining reliable communication is very important [1], [2]. To reliably carry a message from a speaking user to a listening user in such a severe communication scenario, adopting a non-acoustic sensor such as a bone-conduction (BC) microphone is a good choice because the BC microphone is immune to ambient noise by the virtue of its physical sensing characteristic [3]–[5].

There are many approaches to adopt BC microphones for speech enhancement. For example, a single-channel BC microphone-only system uses a speaker-dependent or common equalization (EQ) filter to adopt the characteristics of the BC speech to normal air-conduction (AC) speech [6]–[8]. The role of the EQ filter is to emphasize high-frequency components because the BC sensor typically introduces a low-pass filter effect. However, it does not generate a natural enhanced speech output because it is not easy to reconstruct high-frequency components using a generalized compensation rule. Therefore, research on speech enhancement using a BC microphone commonly adopt multi-channel approaches that utilize both AC and BC microphones. For example, a BC microphone was used as a supplementary sensor for an AC microphone-based system. By using the BC microphone for detecting voice active or inactive regions, they tried to improve the performance of the noise power spectral density (PSD) estimation [9]–[11]. Although these approaches showed some performance improvement, they did not provide how much gain could be achieved for speech enhancement in a highly noisy environment.

Srinivasan and Kechichian quantified the usefulness of a BC microphone using information-theoretic measurements [12]. Specifically, they measured the mutual information (MI) between signals obtained from AC and BC microphones. Although they provided a theoretical guideline, there has been no clear explanation on what is the best strategy of utilizing the sensors for speech enhancement. In other words, it is still unclear how to combine AC and BC microphone input signals in the best way.

Most speech enhancement algorithms such as the Wiener filter or the minimum mean square error (MMSE) based estimator utilize *a priori* signal-to-noise ratio (SNR) and/or *a posteriori* SNR for deriving a multiplicative gain in each frequency bin. Once the noise PSD is estimated by any of these approaches, an *a posteriori* SNR can be computed very easily because its desired signal component is the acquired signal itself. However, the *a priori* SNR still requires a method to estimate the PSD of the desired clean speech signal. A maximum-likelihood (ML) based or a decision-directed (DD) approach are typical methods to estimate the *a priori* SNR. The DD approach takes a weighted sum of the estimated *a priori* SNR in the previous frame and the ML estimate using the *a posteriori* SNR of the current frame. Yu and Fingscheidt introduced a third entity, i.e., the spatial instantaneous SNR, with three weighting factors in total to achieve a multi-channel AC post-filter design [13]. To enhance the robustness of estimation accuracy, the weights are determined by considering the uncertainty of the estimator or frame characteristics. Since the assumption of uncorrelated speech and noise components is not valid as the amount of noise increases, it is still a challenging issue how to correctly estimate the *a priori* SNR under highly noisy conditions.

Note that the *a priori* SNR estimation should not introduce any tracking delay in high SNR conditions and under- or over-

Manuscript received December 02, 2014; revised May 04, 2015; accepted June 08, 2015. Date of publication June 16, 2015; date of current version August 05, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Christensen.

H. S. Shin and H.-G. Kang are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: signal@dsp.yonsei.ac.kr; hgkang@yonsei.ac.kr).

T. Fingscheidt is with the Institute for Communications Technology, Technische Universität Braunschweig, D-38160 Braunschweig, Germany (e-mail: t.fingscheidt@tu-bs.de).

Digital Object Identifier 10.1109/TASLP.2015.2446202

estimation in low SNR conditions. It is reported that the estimation error of the *a priori* SNR in low SNR conditions is more sensitive than in high SNR conditions [14]. Thus, *a priori* SNR estimation in low SNR conditions needs to be designed carefully. To accurately estimate the *a priori* SNR, modified DD estimation approaches and their multi-channel extension cases were introduced [13], [15], [16].

The main idea of the proposed algorithm is to apply the equalized BC speech to the *a priori* SNR estimation process in a DD framework. For *a priori* SNR estimation, we combine three types of estimates, i.e., the estimate from the previous frame, the ML estimate based on the *a posteriori* SNR, and the instantaneous estimate obtained by the equalized BC speech, with their respective weighting factors. In other words, the proposed *a priori* SNR estimator utilizes the strength of the ML estimator in high SNR conditions and the instantaneous BC-based estimator in low SNR conditions. Note that it does not need any training process to generate equalized BC speech. The advantage of the proposed approach is that its weighting factor is adaptively updated based on the variation of environments and the utility of AC and BC microphones. The weighting factors are controlled to have a slow variation of *a priori* SNR in noise-only and steady-state regions, but to allow for a rapid adjustment once the environment changes abruptly.

The performance of the proposed approach is evaluated by implementing it into the Wiener gain estimator which consists of only *a priori* SNR estimator. To evaluate the speech quality of the enhanced speech signal, two perception-based measurements, i.e., the frequency-weighted segmental SNR and the weighted spectral slope, are calculated. Subjective listening tests confirm the superiority of the proposed algorithm to conventional methods in various noise environments.

The remainder of this paper is organized as follows. Section II explains the relationship between the AC and BC channels and the fundamental concept of BC speech enhancement processing. In Section III, we analyze conventional noise reduction and explain the importance of accuracy of *a priori* SNR estimation in low SNR. Section IV describes the proposed *a priori* SNR estimation algorithm using AC and BC microphones. Section V presents the system setup for simulation. Simulation results to the proposed algorithm and analysis by objective and subjective quality measures comparing the proposed method with some of the existing algorithms follow in Section VI. We finalize this paper in Section VII giving some conclusions.

II. Speech Enhancement with AC and BC Microphone

Assuming the additive noise model for each microphone channel, the short-time spectra of AC and BC microphone inputs are defined as follows:

$$Y_i(\ell,k) = S_i(\ell,k) + N_i(\ell,k), \quad i \in \{a,b\}, \tag{1}$$

where $Y_i(\ell,k)$, $S_i(\ell,k)$, and $N_i(\ell,k)$ denote the short-time spectrum of noisy speech, clean speech, and the noise signal from the AC ($i = a$) and the BC ($i = b$) channel, respectively. In addition, ℓ and k denote frame and frequency index, respectively.

The functional effectiveness of using each microphone varies depending on the status of operational environment. For example, the role of the AC signal is more important as the power of the background noise becomes smaller. The BC input is very useful in low SNR environments since its low frequency speech components are similar to the desired clean speech. Recall that the amount of background noise is negligible in the BC microphone signal because the BC microphone mainly picks up the vibration of human bones. However, the BC speech has muffled characteristic due to the attenuation of its high frequency components [17].

The clean desired signal can be estimated by fully utilizing the signals from the AC and the BC microphones such as:

$$\widehat{S_a}(\ell,k) = \delta_a H_a(\ell,k) Y_a(\ell,k) + \delta_b H_b(\ell,k) S_b(\ell,k), \tag{2}$$

where δ_a and δ_b represent the contribution of each microphone to the enhanced signal. For instance, $\delta_a = 1$ and $\delta_b = 0$ becomes typical AC-based single-channel enhancement, and $\delta_a = 0$ and $\delta_b = 1$ represents BC-based enhancement with an equalization process. In other words, $H_b(\ell,k)$ represents a time-varying equalization (EQ) filter to restore the original characteristic of normal speech. Note that the characteristic of filter $H_b(\ell,k)$ varies by the type of input signals such as phonetic- and speaker-specific differences. Various approaches have been proposed to design the EQ filter [6], [7], [17]. $H_a(\ell,k)$ represents a gain function of the conventional AC-based speech enhancement algorithms such as Wiener filter, spectral subtraction, and minimum mean square error (MMSE) based amplitude estimator, log-spectral amplitude estimator (LSA), and so on [18], [19]. The gain function is typically related to the SNR of each frequency bin.

Since each microphone has its own advantage, it is possible to further improve the performance of the overall system if all information from the observed AC and BC speech signals are combined efficiently. We can control the contribution, δ_b and δ_a, as in [20]. Another method is to make $H_a(\ell,k)$ a function of $H_b(\ell,k)$, as in [21]. In this paper, we will incorporate $H_b(\ell,k)$ into the process of estimating the *a priori* SNR, which is very important to determine the gain function $H_a(\ell,k)$ of the noise reduction framework.

III. Analysis of Noise Reduction Using an AC Microphone

In conventional AC-based noise reduction systems, the enhanced output is obtained by multiplying the noise reduction gain, $G(\ell,k)$, to the noisy AC speech, $Y_a(\ell,k)$:

$$\widehat{S_a}(\ell,k) = G(\ell,k) Y_a(\ell,k). \tag{3}$$

The noise reduction gain, $G(\ell,k)$, can be derived from the function of the *a priori* SNR, $\xi_a(\ell,k)$, and/or the *a posteriori* SNR, $\gamma_a(\ell,k)$:

$$G(\ell,k) = \mathcal{F}\left[\xi_a(\ell,k), \gamma_a(\ell,k)\right], \tag{4}$$

$$\xi_a(\ell,k) \equiv \frac{\mathbb{E}\left\{|S_a(\ell,k)|^2\right\}}{\mathbb{E}\left\{|N_a(\ell,k)|^2\right\}}, \tag{5}$$

$$\gamma_a(\ell,k) \equiv \frac{|Y_a(\ell,k)|^2}{\mathbb{E}\left\{|N_a(\ell,k)|^2\right\}}, \tag{6}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator.

Among many speech enhancement algorithms, we focus on the Wiener gain estimator because it only requires the estimated value of the *a priori* SNR, $\xi_a(\ell,k)$. Note that, the proposed algorithm can be easily extended to other gain estimators. The Wiener gain using the AC microphone is obtained as follows:

$$G(\ell,k)=\frac{\widehat{\xi_a}(\ell,k)}{1+\widehat{\xi_a}(\ell,k)}, \tag{7}$$

where $\widehat{\xi_a}(\ell,k)$ denotes the estimated *a priori* SNR value.

A. Decision-Directed (DD) Approach

The *a priori* SNR can be estimated by a decision-directed (DD) approach, i.e., a weighted sum of *a priori* SNR obtained from the previous frame by applying a noise reduction process and the one by a simple variation of *a posteriori* SNR, namely maximum-likelihood (ML) estimate [18], [22].

$$\widehat{\xi_a}(\ell,k)=\alpha_1(\ell,k)\widehat{\xi_a^{(1)}}(\ell-1,k)+\alpha_2(\ell,k)\widehat{\xi_a^{(2)}}(\ell,k), \tag{8}$$

where

$$\widehat{\xi_a^{(1)}}(\ell-1,k)=\frac{|G(\ell-1,k)\cdot Y_a(\ell-1,k)|^2}{\widehat{\sigma_{N_a}^2}(\ell-1,k)}, \tag{9}$$

$$\widehat{\xi_a^{(2)}}(\ell,k)=\mathbb{P}\left[\widehat{\gamma_a}(\ell,k)-1\right]=\mathbb{P}\left[\frac{|Y_a(\ell,k)|^2}{\widehat{\sigma_{N_a}^2}(\ell,k)}-1\right], \tag{10}$$

with $\widehat{\sigma_{N_a}^2}(\ell,k)$ being the estimated noise PSD. $\widehat{\gamma_a}(\ell,k)$ denotes the estimated *a posteriori* SNR. $\alpha_i(\ell,k)$ for $i=1,2$ are smoothing factors and $\mathbb{P}[\cdot]$ denotes the half-wave rectification operator. Please note that $\alpha_1(\ell,k)+\alpha_2(\ell,k)=1$. $\widehat{\xi_a^{(1)}}(\ell-1,k)$ denotes the estimated *a priori* SNR obtained by the estimated speech spectral component and the noise PSD in the previous analysis frame. $\widehat{\xi_a^{(2)}}(\ell,k)$ denotes the ML estimate of the current frame [23].

The conventional DD algorithm employs a larger weight to $\widehat{\xi_a^{(1)}}(\ell-1,k)$ and a smaller weight to $\widehat{\xi_a^{(2)}}(\ell,k)$ to slowly update the value. To control the trade-off between musical noise and speech distortion, $\alpha_1(\ell,k)$ is typically chosen in the range of 0.92 to 0.98. Since this causes strong temporal smoothing, it is not suitable for tracking rapid signal variation in transition regions. Not that the conventional DD approach results in one frame tracking delay in high SNR conditions [24]. Since it is not possible to solve the tracking delay problem using the constant weighting factor, it is necessary to adjust the weighting factor depending on the frame and frequency characteristics. To account for abrupt changes of the speech spectral amplitudes, a self-adaptive smoothing coefficient was proposed using the MMSE criterion explained in the next subsection [25].

B. Weighting Factors

The optimum weighting factors can be derived by the MMSE criterion, $\mathfrak{D}$, as follows:

$$\mathfrak{D}=\mathbb{E}\left[\left\{\xi_{\mathfrak{a}}(\ell,\mathfrak{k})-\widehat{\xi_{\mathfrak{a}}}(\ell,\mathfrak{k})\right\}^2\right]. \tag{11}$$

By substituting Eq. (8) into Eq. (11), the optimum weighting factor of $\alpha_1(\ell,k)$ is obtained by setting the partial derivative to zero, $\partial\mathfrak{D}/\partial\alpha_1=0$:

$$\alpha_1(\ell,k)=\frac{1}{1+\left(\frac{\xi_a(\ell,k)-\widehat{\xi_a^{(1)}}(\ell-1,k)}{\xi_a(\ell,k)+1}\right)^2}. \tag{12}$$

To derive the factor, the relation from the definition of the fourth moment $\mathbb{E}\{|S_a(\ell,k)|^4\}/\mathbb{E}|\{|N_a(\ell,k)|^4\}=2\xi_a^2(\ell,k)$ and $\mathbb{E}\{(\gamma_a(\ell,k)-1)^2\}=2\xi_a^2(\ell,k)+2\xi_a(\ell,k)+1$ are substituted [25].

If the error between $\xi_a(\ell,k)$ and $\widehat{\xi_a^{(1)}}(\ell-1,k)$ is small, $\alpha_1(\ell,k)$ approaches one. As the error increases and $\xi_a(\ell,k)$ decreases, $\alpha_1(\ell,k)$ gets close to zero. Please note that $\xi_a(\ell,k)$ is unknown in Eq. (12), therefore it is substituted by the ML estimate $\widehat{\xi_a^{(2)}}(\ell,k)$ [25]. This is a reasonable alternative because the ML estimate in high SNR is reliable [15]. However, it introduces a large error in low SNR conditions due to the assumption of uncorrelated speech and noise explained in Appendix A.

Note that the *a priori* SNR should be computed by utilizing the information of the current frame in order to avoid any delay. In the framework of using both AC and BC microphones, it can be implemented by using the equalized BC speech of the current frame. Since the BC speech is immune to background noise, the new *a priori* SNR estimate with the equalized BC speech can provide complementary information to the conventional DD approach, especially in low SNR conditions. In the next section, we propose an efficient method to enhance the accuracy of the *a priori* SNR estimate.

IV. Proposed *a priori* SNR Estimation

In order to utilize the advantage of using the enhanced BC speech under the framework of the conventional DD approach, this paper proposes a method to combine three types of *a priori* SNR estimates that have time-frequency varying weights as follows:[1]

$$\begin{aligned}\widehat{\xi_a}(\ell,k)=&\alpha_1(\ell,k)\widehat{\xi_a^{(1)}}(\ell-1,k)\\&+\alpha_2(\ell,k)\widehat{\xi_a^{(2)}}(\ell,k)+\alpha_3(\ell,k)\widehat{\xi_a^{(3)}}(\ell,k),\end{aligned} \tag{13}$$

where

$$\widehat{\xi_a^{(3)}}(\ell,k)=\frac{|H_b(\ell,k)\cdot S_b(\ell,k)|^2}{\widehat{\sigma_{N_a}^2}(\ell,k)}. \tag{14}$$

The values $\alpha_i(\ell,k)$ for $i\in\{1,2,3\}$ in Eq. (13) denote the weighting factors and $\sum_{i=1}^{3}\alpha_i(\ell,k)=1$. $\widehat{\xi_a^{(1)}}(\ell-1,k)$ and $\widehat{\xi_a^{(2)}}(\ell,k)$ are presented in Eq. (9) and (10), respectively. $\widehat{\xi_a^{(3)}}(\ell,k)$ describes the instantaneous *a priori* SNR estimate obtained by the equalized BC speech. $H_b(\ell,k)$ represents the EQ filter, which is a time-varying filter to restore the original

[1]In this section, $\mathfrak{D}$ denotes the MMSE criterion explained in Eq. (11). $\mathbb{E}\{\cdot\}$ denotes the expectation operator. $\widehat{\xi_a}(\ell,k)$ in Eq. (13) represents the estimated *a priori* SNR value. All α s ($\alpha_i(\ell,k)$ for $i\in\{1,2,3\}$, $\widetilde{\alpha}$ and $\widehat{\alpha}$ in Eq. (20), (21), and (22)) relate with the weighting factors. An amount of utility (AoU), $\rho(\ell,k)$, can be obtained by and $\bar{\rho}(\ell,k)$ in Eq. (18) as a function of a BC-based local SNR (bSNR), $\eta(\ell,k)$ in Eq. (17).

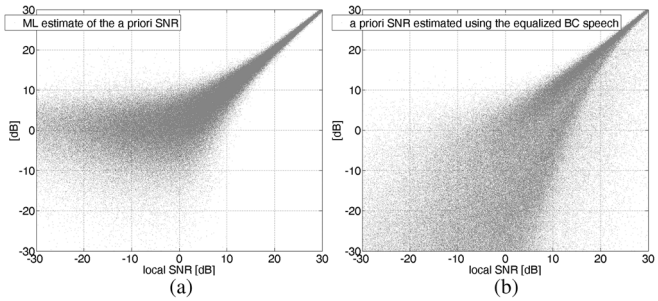

Fig. 1. Distribution of $\widehat{\xi_a^{(2)}}(\ell,k)$ (*left*) and $\widehat{\xi_a^{(3)}}(\ell,k)$ (*right*) as a function of the local time-frequency SNR (*local* SNR). The noisy signal is comprised of a voiced speech segment embedded in car noise (C1) at 0 dB input SNR. Noise PSD is estimated using the minimum statistics noise estimation (MSNE) algorithm [26].

characteristic of normal speech. Thus, the numerator of Eq. (14) can be regarded as $|H_b(\ell,k)\cdot S_b(\ell,k)|^2 \approx |S_a(\ell,k)|^2$.

To understand the characteristics of the proposed estimate $\widehat{\xi_a^{(3)}}(\ell,k)$, we examine its distribution w.r.t. local time-frequency SNRs (*local* SNRs). Please note that the *local* SNR depicted in Fig. 1 denotes an *a priori* SNR, which is computed by the known signal and noise components, $|S_a(\ell,k)|^2/|N_a(\ell,k)|^2$. The distribution of the ML estimate of the *a priori* SNR, $\widehat{\xi_a^{(2)}}(\ell,k)$ (*left panel*), and the instantaneous *a priori* SNR using the equalized BC speech, $\widehat{\xi_a^{(3)}}(\ell,k)$ (*right panel*), are illustrated in Fig. 1. The gray dots denote the estimated *a priori* SNRs in voice active frames.

In low *local* SNR frames, $\widehat{\xi_a^{(3)}}(\ell,k)$ is under-estimated as shown in Fig. 1(b), whereas $\widehat{\xi_a^{(2)}}(\ell,k)$ is mostly over-estimated. Therefore, we may conclude that $\widehat{\xi_a^{(2)}}(\ell,k)$ and $\widehat{\xi_a^{(3)}}(\ell,k)$ complement each other when we estimate *a priori* SNRs. This trend is consistent even if we change the type of noise PSD estimator to the minima controlled recursive averaging (MCRA) algorithm.

Since the behavior of the *a priori* SNR is controlled by the weighting factors, it is very important how to choose the weighting factors. We derived the weighting factors to balance these three types of *a priori* SNR estimates. To prevent speech distortion in onset regions, a much smaller weighting factor should be used for $\widehat{\xi_a^{(1)}}(\ell-1,k)$,

In speech absent and silence regions, it is appropriate to use a higher value of $\alpha_1(\ell,k)$ because this reduces musical noise artifacts. If the error between the reference SNR and the estimator at the previous frame increases in low SNR cases, it needs a fast update with the value estimated by the information of the current frame. Therefore, the weighting factors should be designed to push $\alpha_1(\ell,k)$ close to zero.

The proposed *a priori* SNR estimate utilizes time-frequency varying weighting factors to combine the three types of estimates to keep the balance between speech quality improvement and noise reduction.

A. Derivation of Weighting Factors

There are several combinations of taking two from three variables when we deal with an AC and a BC microphone: 1) $\widehat{\xi_a^{(1)}}(\ell-1,k)$ and $\widehat{\xi_a^{(2)}}(\ell,k)$, 2) $\widehat{\xi_a^{(1)}}(\ell-1,k)$ and $\widehat{\xi_a^{(3)}}(\ell,k)$, and 3) $\widehat{\xi_a^{(2)}}(\ell,k)$ and $\widehat{\xi_a^{(3)}}(\ell,k)$. For simplicity, a vector representation such as $\boldsymbol{\alpha}_{\ell,k} = (\alpha_1(\ell,k),\alpha_2(\ell,k),\alpha_3(\ell,k))$ is used in this Section.

1) Case of $\widehat{\xi_a^{(1)}}$ and $\widehat{\xi_a^{(2)}}$: In Section III-B, we explained the conventional DD approach that does not include the BC-based *a priori* estimate. In this case, the unknown value of $\xi_a(\ell,k)$ in Eq. (12) is substituted for $\widehat{\xi_a^{(3)}}(\ell,k)$ only to calculate the weighting factors. In next sub-sections, we describe the other cases.

2) Case of $\widehat{\xi_a^{(1)}}$ and $\widehat{\xi_a^{(3)}}$: Assuming that $\widehat{\xi_a^{(2)}}(\ell,k)$ is disregarded, the *a priori* SNR is composed of a weighted sum of *a priori* SNR obtained in the previous frame and the one with the equalized BC speech. To find the optimum value of the weighting factor, we substitute Eq. (13) with $\alpha_2(\ell,k)=0$ into Eq. (11). The optimal weighting factor is derived by setting $\partial\mathfrak{D}/\partial\alpha_1 = \mathfrak{o}$:

$$\alpha_1(\ell,k) = \frac{1}{1+\left(\frac{\xi_a(\ell,k)-\widehat{\xi_a^{(1)}}(\ell-1,k)}{\xi_a(\ell,k)}\right)^2}, \tag{15}$$

where $\boldsymbol{\alpha}_{\ell,k} = (\alpha_1(\ell,k),0,1-\alpha_1(\ell,k))$. Eq. (15) represents a temporal smoothing process to prevent musical noise.

If the temporal difference between $\xi_a(\ell,k)$ and $\widehat{\xi_a^{(1)}}(\ell-1,k)$ is small, $\alpha_1(\ell,k)$ in Eq. (15) approaches to one. As the error increases and $\xi_a(\ell,k)$ decreases, $\alpha_1(\ell,k)$ becomes close to zero. Since $\xi_a(\ell,k)$ is unknown, we substitute $\widehat{\xi_a^{(3)}}(\ell,k)$ for $\xi_a(\ell,k)$ in Eq. (15). Although $\widehat{\xi_a^{(2)}}(\ell,k)$ is dispersed, it is a reasonable alternative since the equalized BC-speech-based instantaneous SNR provides nice results due to independency of the *a posteriori* SNR term. Furthermore, $\widehat{\xi_a^{(3)}}(\ell,k)$ results in under-estimation of $\xi_a(\ell,k)$ in low SNR.

3) Case of $\widehat{\xi_a^{(2)}}$ and $\widehat{\xi_a^{(3)}}$: Let's consider the case when the estimated *a priori* SNR obtained in the previous frame $\widehat{\xi_a^{(1)}}(\ell-1,k)$ is disregarded. Since we assume no dependency on the past values in this case, the processing is done by two *a priori* SNR estimates of the current frame. It enables us to reduce the tracking delay, thus to follow abrupt changes in transient regions. By incorporating $\alpha_1(\ell,k)=0$ into Eq. (13), it is substituted into Eq. (11). In this case, the optimal weighting factor is derived by taking $\partial\mathfrak{D}/\partial\alpha_2 = \mathfrak{o}$:

$$\boldsymbol{\alpha}_{\ell,k} = (0,0,1). \tag{16}$$

Thus, it follows the *local* SNR directly obtained by the enhanced BC speech component. It is not surprising that the ML estimate of *a priori* SNR is disregarded because of the assumption on $\mathbb{E}[\widehat{\xi_a^{(3)}}(\ell,k)] \equiv \xi_a(\ell,k)$. The result shows that, in the MMSE sense, the BC-only approach can be applied if $\widehat{\xi_a^{(3)}}(\ell,k)$ is highly reliable.

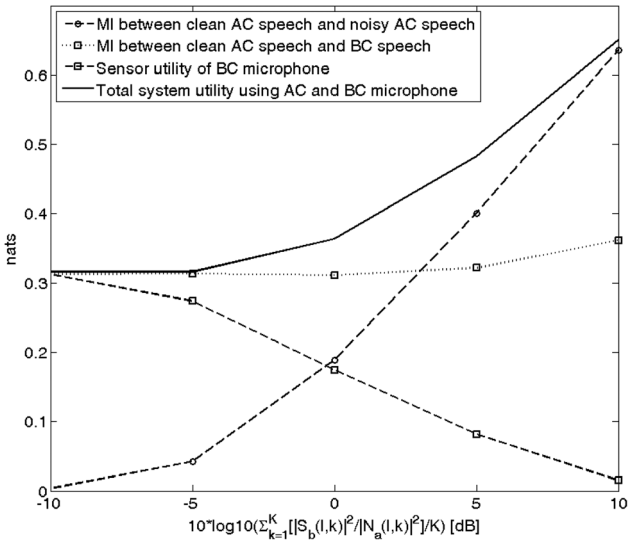

Fig. 2. Example of utility measures as a function of a frequency-averaged BC-based local SNR. (K denotes the total number of frequency bins. The amount of information is measured in nats.).

In the low frequency region, such an assumption is reasonable since $\widehat{\xi_a^{(3)}}(\ell,k)$ is sufficiently reliable. Thus, the BC stand-alone system is a good alternative. Although the recovery of speech components by only using $\widehat{\xi_a^{(3)}}(\ell,k)$ is not perfect in high frequency regions, instantaneous *a priori* SNR estimation leads to no excessive amplification of residual noise.

4) All Cases: We will now introduce weighting factors embracing all cases mentioned above. To preserve speech components, the weighting factors should be designed to control $\widehat{\xi_a^{(1)}}(\ell-1,k)$, $\widehat{\xi_a^{(2)}}(\ell,k)$, and $\widehat{\xi_a^{(3)}}(\ell,k)$ by introducing reasonable criteria. In this paper, a utility measure is adopted as a criterion for the control mechanism of the weighting factors.

B. Determination of Weighting Factors by a Utility Measure

It is important to design an efficient evaluation criterion for determining the weighting factors of the three types of *a priori* SNR estimates. This paper presents the amount of information that can be achieved by combining AC and BC microphones. The way of quantifying the utility of a BC microphone using an information theoretic measurement was introduced in [12]. From the measured mutual information (MI), we can conjecture the best scenario of using the BC microphone for speech enhancement. The sensor and total utility measures were estimated using the entropy estimation based on a Gaussian mixture model (GMM) [27].

Fig. 2 depicts the sensor and total system utility measures following the method given in [12]. White noise with varying power was added to the recorded clean AC speech signals. Mel-frequency cepstral coefficients (MFCCs) were used as features. The x-axis and y-axis represent a frequency-averaged BC-based local SNR in [dB] and the amount of information which is measured in nats since natural logarithms are used, respectively. The sensor utility of the BC microphone is defined by the conditional MI between clean AC speech and noisy BC speech given noisy AC speech. Also, the total system utility using AC and BC microphone is defined by the sum of the sensor utility of the BC microphone and the MI between clean AC speech and noisy AC speech [12]. The BC signal that contains the partial information of the clean AC speech signal is useful for the case of low SNR frames, but its utility value (*dashed line with square*) decreases as the SNR increases because the clean AC speech signal contains most of the information in high SNR cases (see Fig. 2). The BC microphone contributes to the estimation of the clean AC speech before the total system utility (*solid line*) coincides with MI between the clean AC speech and the noisy AC speech (*dashed line with circle*). In Fig. 2, the impact of the BC microphone becomes higher if the SNR of the frame is lower than 10 dB. Although the utility measures vary depending on SNRs, it is hard to apply the utility measure of the BC microphone directly for real-time systems.

We define a BC-based local SNR (bSNR) by the ratio between the speech PSD of the BC microphone and the noise PSD of the AC microphone as a replacement of the utility measure:

$$\eta(\ell,k) \equiv \frac{|S_b(\ell,k)|^2}{\mathbb{E}\left\{|N_a(\ell,k)|^2\right\}}. \tag{17}$$

The weighting factors can be adaptively controlled depending on the bSNR. A higher utility of the BC microphone is guaranteed when the local SNR is low. Thus, it is reasonable to assign higher weighting factors to the *a priori* SNR derived by the BC speech in low bSNRs. The speech-dominant region can be also detected by utilizing $S_b(l,k)$.

Since the bSNR, $\eta(\ell,k)$, has virtually infinite dynamic range, it is required to design a mapping function to the range of 0 to 1. We introduce a hyperbolic tangent function to measure utility as follows:

$$\bar{\rho}(\ell,k) = \tanh\left(-\eta(\ell,k)\right) + 1. \tag{18}$$

This measure is designed to assign higher values to frames with lower bSNRs. Since $\bar{\rho}(\ell,k)$ has a limitation of accuracy in the high frequency region, it has weakness in detecting spectral harmonic components that show a high value of local SNR.

The extension of $\bar{\rho}(\ell,k)$ to the high frequency region is solved by a spectral folding technique. We take the minimum value of $\bar{\rho}(\ell,k)$ and the mirror image of $\bar{\rho}(\ell,k)$ as an amount of utility (AoU), $\rho(\ell,k)$, to increase the weight of the ML estimate to consider harmonic spectral components at higher frequencies. The AoU can be obtained as:

$$\rho(\ell,k) = \min\left(\bar{\rho}(\ell,k), \bar{\rho}(\ell,K-k)\right), \tag{19}$$

where K denotes the number of frequency bins. The AoU, $\rho(\ell,k)$, provides a guideline to determine which method, the ML estimate or the instantaneous SNR based on the BC information, is reliable in what conditions.

Finally, the weighting factor given in Section IV-A is modified by adopting the proposed AoU, $\rho(\ell,k)$. A vector representation $\boldsymbol{\alpha}_{\ell,k} = (\alpha_1(\ell,k), \alpha_2(\ell,k), \alpha_3(\ell,k))$ is used. We omit the indices ℓ, k in Eq. (20) for notational convenience.

$$\boldsymbol{\alpha}_{\ell,k} = \left((1-\rho)\,\widetilde{\alpha} + \rho\widehat{\alpha}, (1-\rho)\,(1-\widetilde{\alpha}), \rho\,(1-\widehat{\alpha})\right), \tag{20}$$

where

$$\widetilde{\alpha}(\ell,k)=\frac{1}{1+\left(\frac{\xi_a(\ell,k)-\widehat{\xi_a^{(1)}}(\ell-1,k)}{\xi_a(\ell,k)+1}\right)^2}, \tag{21}$$

$$\widehat{\alpha}(\ell,k)=\frac{1}{1+\left(\frac{\xi_a(\ell,k)-\widehat{\xi_a^{(1)}}(\ell-1,k)}{\xi_a(\ell,k)}\right)^2}, \tag{22}$$

and $\sum_{i=1}^{3}\alpha_i(\ell,k)=1$. We make full use of the ML-based and the equalized BC-speech-based estimates by adopting the AoU. Please note that Eq. (20) includes all the cases: Case *1)* ($\rho(\ell,k)=0$); Case *2)* ($\rho(\ell,k)=1$); and Case *3)* ($\rho(\ell,k)=1$ and $\widehat{\alpha}(\ell,k)=0$).

Since $\xi_a(\ell,k)$ given in Eq. (21) and (22) is unknown, we need to estimate it. In high SNR cases, the ML estimate can be still used. However, in low SNR cases, it is a better choice of using $\widehat{\xi_a^{(3)}}(\ell,k)$ rather than $\widehat{\xi_a^{(2)}}(\ell,k)$. Thus, we substitute $\xi_{a,p}(\ell,k)$ in Eq. (23) for $\xi_a(\ell,k)$ in Eq. (21) and (22), which is defined as

$$\xi_{a,p}(\ell,k)=\begin{cases}\widehat{\xi_a^{(2)}}(\ell,k), & \eta_{dB}(\ell)>\eta_{thr},\\ \widehat{\xi_a^{(3)}}(\ell,k), & \text{otherwise},\end{cases} \tag{23}$$

where $\eta_{dB}(\ell)=10\log 10(\frac{1}{K}\sum_{k=1}^{K}\eta(\ell,k))$. η_{thr} is the threshold that represents the point where the amount of the MI between the clean speech and the noisy AC speech crosses that of the MI between the clean speech and the BC speech (see Fig. 2).

C. Analysis of the Proposed Weighting Factors

The weighting factors vary depending on the values of the temporal difference of *a priori* SNR, $\Delta(\ell,k)=\xi_a(\ell,k)-\widehat{\xi_a^{(1)}}(\ell-1,k)$. For smaller Δ, the effect of $\alpha_2(\ell,k)$ and $\alpha_3(\ell,k)$ reduces while $\alpha_1(\ell,k)$ increases.

An example dependency of the weighting factors for different AoU, the sensor utility measure of the BC microphone, is illustrated in Fig. 3 when the temporal difference is high ($\Delta=10$). The weighting factor curves are depicted for an AoU of $\rho=0$ in Fig. 3(a), and $\rho=0.8$ in Fig. 3(b). The values $\alpha_1(\ell,k)$, $\alpha_2(\ell,k)$, and $\alpha_3(\ell,k)$ are depicted as solid line with asterisks, with circles, and with squares, respectively.

When the AoU is equal to zero ($\rho=0$), it becomes the conventional DD approach as shown in Fig. 3(a). Since $\alpha_2(\ell,k)$ increases as Δ increases, it enables the *a priori* SNR estimate to follow abrupt changes promptly. For the DD approach, the dominant factor in low SNR is $\alpha_2(\ell,k)$. Thus, it increases the residual noise components in low SNR conditions. To improve the performance in highly noisy conditions, the weighting factors should be changed to allow for a higher value of $\alpha_3(\ell,k)$. This is why the AoU should be adopted to control the weighting factors.

Depending on the AoU, we can control the contribution of three *a priori* SNR estimates by the weighting factors appropriately. The higher AoU emphasizes $\alpha_3(\ell,k)$ in low SNR to increase the effect of the robust $\widehat{\xi_a^{(3)}}(\ell,k)$. When the AoU is high in low SNR, $\alpha_3(\ell,k)$ plays a dominant role in supporting $\alpha_2(\ell,k)$ or $\alpha_1(\ell,k)$ with large Δ.

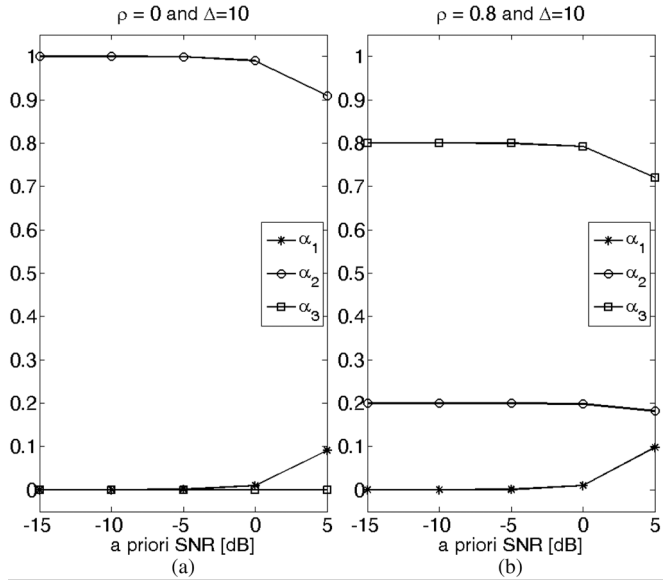

Fig. 3. Weighting factors in Eq. (20) as a function of the *a priori* SNR for different values of $\Delta(\ell,k)=\xi_a(\ell,k)-\widehat{\xi_a^{(1)}}(\ell-1,k)$. The curves are obtained using a fixed value of (a) $\rho=0$, (b) $\rho=0.8$.

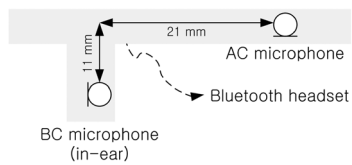

Fig. 4. Position of AC and BC microphones embedded in the bluetooth headset.

TABLE I
NOISE CONDITIONS FOR SAMPLE RECORDING (THE ENGINE IS ON AND ALL CAR WINDOWS ARE CLOSED FOR EVERY CONDITION.)

Noise Condition	Air Conditioner	Speed
C1	Off	0 km/h
C2	Off	50 km/h
C3	Off	120 km/h
C4	On	50 km/h
C5	On	120 km/h

D. Equalization Filter Estimation

The best way to design an EQ filter is to use the same speaker's clean AC speech sentence of the same phrase as the clean BC speech sentence [7]. If the EQ filter and the spectrum of clean BC speech were known, perfect reconstruction would be possible. However, the EQ filter and the BC clean speech spectrum have to be estimated in practical applications.

To estimate the EQ filter designed by the minimizing mean square error, we need the clean speech spectrum of each channel. Since the observed signals are noisy speech signals, estimated AC and BC speech signals are used to calculate power spectra in practical situations:

$$\widehat{H_b}(\ell,k)=\frac{\mathrm{P}_{ab}(\ell,k)}{\mathrm{P}_{bb}(\ell,k)}, \tag{24}$$

where $\mathrm{P}_{ab}(\ell,k)$ is the cross power spectrum of between clean AC and BC speech, $\mathrm{P}_{ab}(\ell,k)=\widetilde{S_a^*}(\ell,k)\widetilde{S_b}(\ell,k)$. $\mathrm{P}_{bb}(\ell,k)$ is the power spectrum of clean BC speech,

TABLE II
FREQUENCY-WEIGHTED SEGMENTAL SNR (fwSNRseg) WITH RESPECT TO INPUT SNR (THE BEST PERFORMANCE IS SHOWN IN BOLD. THE UPPER BOUND IS OBTAINED BY SUBSTITUTING THE *LOCAL* SNR FOR THE *A PRIORI* SNR ESTIMATOR WHEN WE CALCULATE THE NOISE REDUCTION GAIN. NUMBERS IN PARENTHESIS INDICATE THE STANDARD DEVIATION.)

Noise Type	SNR	DD approach		Proposed approach				Upper bound
		fixed	*adaptive*	*Case 1*	*Case 2*	*Case 3*	*AoU*	*local* SNR
Mixed noise (Six different noise types taken from NOISEX-92 database [28])	-15 dB	1.77 (0.41)	1.68 (0.30)	1.65 (0.30)	2.33 (0.31)	2.37 (0.26)	**2.44** (0.33)	7.47 (0.28)
	-10 dB	2.34 (0.52)	2.14 (0.37)	2.09 (0.36)	2.82 (0.27)	2.88 (0.25)	**3.24** (0.33)	8.14 (0.32)
	-5 dB	3.16 (0.61)	2.98 (0.46)	2.89 (0.45)	3.47 (0.24)	3.55 (0.25)	**4.23** (0.28)	8.82 (0.43)
	0 dB	4.23 (0.63)	4.23 (0.54)	4.12 (0.52)	4.36 (0.27)	4.47 (0.29)	**5.29** (0.29)	9.63 (0.52)
	5 dB	5.55 (0.53)	5.87 (0.59)	5.77 (0.57)	5.57 (0.31)	5.71 (0.33)	**6.52** (0.27)	10.66 (0.58)
	10 dB	6.96 (0.39)	7.80 (0.61)	7.74 (0.60)	6.96 (0.36)	7.13 (0.34)	**7.98** (0.28)	11.99 (0.64)
	15 dB	8.39 (0.41)	9.97 (0.61)	9.93 (0.61)	8.41 (0.50)	8.61 (0.48)	**9.74** (0.31)	13.69 (0.68)
Automobile noise (Recorded noise under five conditions in Table I)	-15 dB	4.13 (0.15)	3.67 (0.41)	3.59 (0.42)	4.28 (0.30)	4.42 (0.30)	**5.21** (0.35)	8.45 (0.49)
	-10 dB	5.17 (0.30)	4.69 (0.39)	4.60 (0.40)	5.18 (0.45)	5.39 (0.46)	**6.30** (0.43)	9.74 (0.56)
	-5 dB	6.55 (0.48)	6.08 (0.39)	6.01 (0.40)	6.54 (0.62)	6.74 (0.61)	**7.65** (0.48)	11.18 (0.63)
	0 dB	8.28 (0.64)	7.82 (0.38)	7.77 (0.39)	8.19 (0.75)	8.37 (0.73)	**9.25** (0.53)	12.83 (0.69)
	5 dB	10.19 (0.68)	9.88 (0.38)	9.84 (0.38)	10.03 (0.75)	10.15 (0.71)	**11.17** (0.52)	14.69 (0.72)
	10 dB	12.14 (0.68)	12.27 (0.36)	12.23 (0.36)	11.89 (0.68)	11.94 (0.65)	**13.18** (0.53)	16.71 (0.70)
	15 dB	14.01 (0.61)	14.81 (0.37)	14.77 (0.37)	13.58 (0.60)	13.55 (0.58)	**14.94** (0.53)	18.77 (0.61)

TABLE III
WEIGHTED SPECTRAL SLOPE DISTANCE (d_{WSS}) WITH RESPECT TO INPUT SNR (THE BEST PERFORMANCE IS SHOWN IN BOLD. THE LOWER BOUND IS OBTAINED BY SUBSTITUTING THE *LOCAL* SNR FOR THE *A PRIORI* SNR ESTIMATOR WHEN WE CALCULATE THE NOISE REDUCTION GAIN. NUMBERS IN PARENTHESIS INDICATE THE STANDARD DEVIATION.)

Noise Type	SNR	DD approach		Proposed approach				Lower bound
		fixed	*adaptive*	*Case 1*	*Case 2*	*Case 3*	*AoU*	*local* SNR
Mixed noise (Six different noise types taken from NOISEX-92 database [28])	-15 dB	165.73 (16.86)	148.78 (15.07)	148.40 (14.96)	120.72 (9.92)	123.30 (9.64)	**118.95** (10.00)	46.66 (3.85)
	-10 dB	153.01 (14.81)	138.43 (12.81)	138.33 (12.63)	109.22 (9.31)	112.01 (8.64)	**102.35** (8.22)	42.55 (2.73)
	-5 dB	137.41 (13.14)	124.64 (10.99)	124.77 (10.83)	95.89 (8.82)	98.40 (8.01)	**85.55** (6.83)	38.22 (1.95)
	0 dB	120.80 (11.45)	109.11 (9.83)	109.18 (9.73)	82.20 (7.80)	84.19 (7.12)	**73.30** (5.79)	34.12 (1.45)
	5 dB	104.98 (10.50)	94.21 (8.91)	94.07 (8.83)	69.53 (6.17)	70.94 (5.68)	**64.60** (4.36)	30.41 (1.39)
	10 dB	91.88 (9.62)	81.65 (7.83)	81.33 (7.71)	59.93 (4.58)	60.90 (4.06)	**57.72** (3.90)	26.99 (1.44)
	15 dB	80.72 (8.41)	71.09 (6.77)	70.72 (6.70)	53.16 (3.49)	53.68 (3.11)	**50.84** (3.35)	23.58 (1.40)
Automobile noise (Recorded noise under five conditions in Table I)	-15 dB	108.86 (7.43)	105.60 (5.56)	105.59 (5.55)	87.54 (6.07)	89.84 (6.83)	**74.91** (6.91)	35.07 (1.60)
	-10 dB	93.27 (7.41)	91.90 (5.24)	91.87 (5.18)	73.32 (6.38)	75.08 (6.80)	**62.75** (5.69)	30.78 (1.64)
	-5 dB	77.54 (6.99)	77.93 (5.00)	77.78 (4.97)	60.15 (6.36)	61.34 (6.64)	**53.84** (4.18)	26.83 (1.60)
	0 dB	64.13 (5.62)	65.07 (4.47)	64.89 (4.51)	49.80 (5.37)	50.65 (5.27)	**47.11** (3.32)	23.20 (1.59)
	5 dB	53.57 (4.09)	54.12 (3.64)	53.96 (3.60)	42.28 (3.26)	43.12 (3.06)	**40.92** (2.35)	19.70 (1.51)
	10 dB	45.50 (3.43)	45.19 (3.10)	45.04 (3.03)	37.61 (1.70)	38.29 (1.55)	**36.17** (1.77)	16.25 (1.38)
	15 dB	38.98 (2.90)	37.74 (2.76)	37.64 (2.71)	34.27 (1.18)	34.77 (1.32)	**32.47** (1.58)	12.95 (1.25)

$P_{bb}(\ell,k) = \widetilde{S_b^*}(\ell,k)\widetilde{S_b}(\ell,k)$. $\widetilde{S_a}(\ell,k)$ and $\widetilde{S_b}(\ell,k)$ are the estimated clean AC and BC speech spectra by pre-processing. The superscript $*$ denotes the complex conjugate. $\widehat{H_b}(\ell,k)$ is the estimated EQ filter which compensates the characteristics between AC and BC speech. It is estimated on a frame-by-frame basis to reflect the phoneme change in each frame. The speaker-dependent EQ filter works as a high-pass filter which amplifies the high frequency components of BC speech. $|\widehat{H_b}(\ell,k)\cdot\widetilde{S_b}(\ell,k)|^2$ is substituted into the numerator of the instantaneous *a priori* SNR estimate in Eq. (14).

V. METHODOLOGY AND SIMULATION SETUP

Recordings from the AC and BC channel were made synchronously with the clean speech signals being spoken from the front passenger seat and the background noises recorded separately in a Volkswagen car. A bluetooth headset embedded with a normal AC microphone and an in-ear type BC microphone were used to acquire speech and noise signals, separately. The positioning of the two microphones are illustrated in Fig. 4. The sampling rate was initially 48 kHz and 16 bits were allocated to each sample of the recorded signal. For simulation, speech and noise signals were down-sampled to 8 kHz. Clean speech recorded in an idle state of the car engine was assumed as the target signal in a quiet environment. Speech samples in English were recorded by two speakers (6 sentences for a male and 7 sentences for a female) with each sample lasting longer than 1 minute. As shown in Table I, noise conditions were varied by different driving speeds and by turning on or off the air condition. Note that speech signals were corrupted by noise from the engine, contact between tires and road, air condition, etc. Alternatively, only the clean speech signals were corrupted by six different noise types (babble, buccaneer, destroyerengine, destroyerops, f16, and factory noise; namely mixed noise in Table II and III) taken from NOISEX-92 database [28].

The speech and noise signals were added in each channel in each simulation condition. The level of the AC and BC speech signals were set to -26 dBov according to ITU-T Recommendation P.56 [29]. The BC noise level was 23 dB lower than the

AC noise level on average. The AC noise signal had been obtained by scaling the recorded AC noise signal w.r.t. to the active speech level to control the input SNR, ranging from -15 to 15 dB in steps of 5 dB. The BC noise signal was calibrated by the same factor obtained from the AC speech and the AC noise signal.

Each analysis frame consisted of 256 samples with 50% overlap. For *a priori* SNR and *a posteriori* SNR estimation, a noise power estimator based on minimum statistics was utilized [29]. The *a priori* SNR flooring was done with the lower limit of $\xi_{min} = -60$ dB, i.e., $\max(\widehat{\xi_a}(\ell,k), \xi_{min})$ was utilized instead of $\widehat{\xi_a}(\ell,k)$ in Eq. (13). η_{thr} was set to $\eta_{thr} = 5$ dB in Eq. (23).

VI. Experimental Results

The estimation accuracy of the proposed method is compared with that of the conventional DD approach. The weighting values of conventional DD approaches in Eq. (8), are set to either the *fixed* value of $\alpha_1(\ell,k) = 0.98$, or the *adaptive* value given in Eq. (12) [25]. The proposed *a priori* SNR in Eq. (13) is controlled by different weighting factors. *Case 1*, *Case 2* and *Case 3* represent the *a priori* SNR estimator described in Section IV-A1), IV-A2) and IV-A3), respectively. *AoU* represents the *a priori* SNR estimator controlled by the AoU as given in Eq. (20).

A. Analysis of the Proposed a Priori SNR and its Weighting Factors

From the observation of the proposed *a prior* SNR estimator and its weighting factors, it turns out that the weighting factors are designed properly as mentioned in Section IV-C. Figs. 5 and 6 illustrate the true and estimated values at 1 kHz and 0.5 kHz with 0 dB input SNR in an automobile noise condition, respectively. In Figs. 5–6, (a) show noisy and clean speech signals in the time domain, (b) depicts the *a priori* SNR estimates in [dB]. In (b), $\alpha_1(\ell,k) = 0.98$ was applied to the conventional DD approach ('conventional') in Eq. (8). The proposed *a priori* estimator with the AoU in Eqs. (13) and (20) is denoted 'proposed'. (c) represents time-varying weights $\alpha_1(\ell,k)$, $\alpha_2(\ell,k)$, and $\alpha_3(\ell,k)$ of the proposed *a priori* SNR estimator with AoU. In (c), the weighting factors are illustrated asterisk for $\alpha_1(\ell,k)$, circle for $\alpha_2(\ell,k)$, and square for $\alpha_3(\ell,k)$, respectively.

Fig. 5 depicts an example output obtained in a speech-only region. With the help of $\alpha_3(\ell,k)$ and $\alpha_2(\ell,k)$, the output of the proposed *a priori* SNR estimate is very close to the *local* SNR. Higher $\alpha_3(\ell,k)$ or $\alpha_2(\ell,k)$ increases the speed of update to quickly adapt to characteristic of current frame. During the transient parts of speech, the proposed method generates large $\alpha_2(\ell,k)$, while $\alpha_1(\ell,k)$ is small.

Fig. 6 depicts an example output obtained in a transient region. In the noise-only frames, $\alpha_3(\ell,k)$ is large and $\alpha_1(\ell,k)$ is small without dependency on $\alpha_2(\ell,k)$. At onset region, the proposed method generates large $\alpha_3(\ell,k)$ with small $\alpha_2(\ell,k)$, while $\alpha_1(\ell,k)$ is close to zero. This allows for a fast response to sudden changes in the signal. It is clearly seen that the proposed approach tracks the true value more quickly. In speech frames, $\alpha_1(\ell,k)$ is large on the right side of Fig. 6 since the error between previous and current frames are small during the steady-state segments of speech.

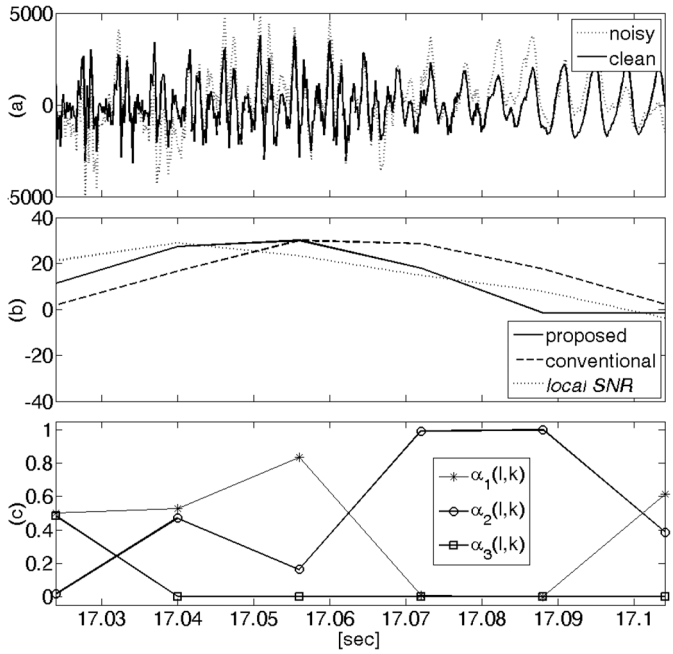

Fig. 5. Automobile noise environment (C5) at 0 dB input SNR at 1.0 kHz: (a) noisy and clean speech signals in the time domain, (b) *a priori* SNR estimates in [dB], and (c) $\alpha_1(\ell,k)$, $\alpha_2(\ell,k)$, and $\alpha_3(\ell,k)$ of the proposed *a priori* SNR estimator with AoU.

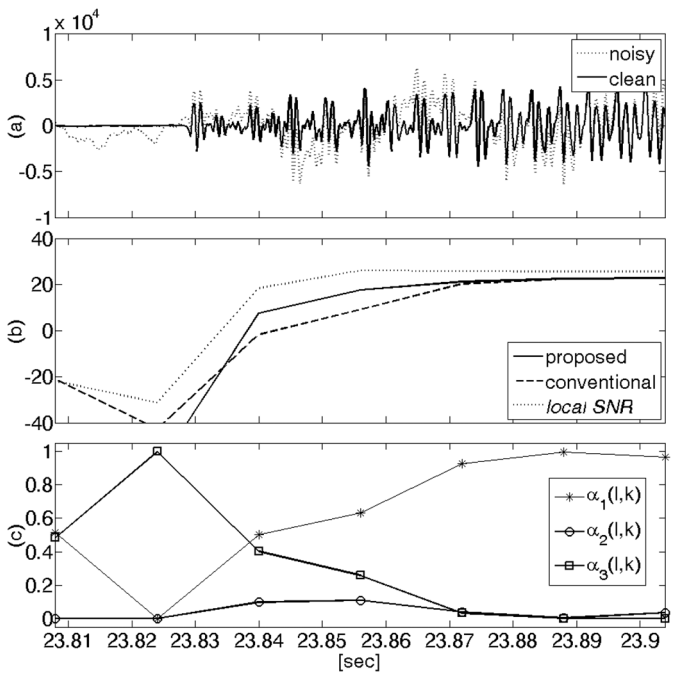

Fig. 6. Automobile noise environment (C5) at 0 dB input SNR at 0.5 kHz: (a) noisy and clean speech signals in the time domain, (b) *a priori* SNR estimates in [dB], and (c) $\alpha_1(\ell,k)$, $\alpha_2(\ell,k)$, and $\alpha_3(\ell,k)$ of the proposed *a priori* SNR estimator with AoU.

B. Evaluation of the Enhanced Speech Quality

Experiments are performed to evaluate the performance of the enhanced speech signal depending on *a priori* SNR estimation methods in various noise types and input SNR conditions.

TABLE IV
SUBJECTIVE TEST RESULT: MOS VALUE WITH RESPECT TO INPUT SNR (THE BEST PERFORMANCE IS SHOWN IN BOLD.)

Noise Type	SNR	DD approach		Proposed approach			
		fixed	*adaptive*	*Case 1*	*Case 2*	*Case 3*	*AoU*
Automobile noise (Recorded noise under five conditions in Table I)	-15 dB	2.14	1.83	1.88	2.43	2.49	**2.58**
	-10 dB	2.63	2.06	2.07	2.84	2.96	**2.98**
	-5 dB	2.88	2.28	2.31	3.11	3.11	**3.18**
	0 dB	3.23	2.67	2.73	3.41	3.39	**3.47**
	5 dB	3.51	3.21	3.26	**3.76**	3.73	3.75
	10 dB	3.82	3.74	3.74	**3.91**	**3.91**	3.88
	15 dB	4.08	**4.10**	4.08	4.03	4.00	4.06

1) Objective Measures: The averaged results are summarized in Tables II and III. Numbers in parenthesis denote the standard deviation. Frequency-weighted segmental SNR (fwSNRseg) and weighted spectral slope distance ($\mathrm{d_{WSS}}$) are used for objective evaluation. fwSNRseg is computed as [30]:

$$\mathrm{fwSNRseg} = \frac{10}{L} \sum_{\ell=1}^{L} \frac{\sum_{j=1}^{K_B} B_j \log_{10}\left[F^2(\ell,j)/\left(F(\ell,j)-\hat{F}(\ell,j)\right)^2\right]}{\sum_{j=1}^{K_B} B_j}, \tag{25}$$

where K_B and L are the total number of bands and frames in the signal, respectively. B_j represents the weight in the j-th frequency band. $F(\ell,j)$ and $\widehat{F}(\ell,j)$ denote the filter-bank amplitude of the clean signal and enhanced signal in the ℓ-th frame and j-th frequency band, respectively. Since fwSNRseg based on the SNR at same time and frequency, it is a measure reflecting perceptual weighting. It has high correlation to subjective listening tests results.

Weighted spectral slope (WSS) distance, $\mathrm{d_{WSS}}$, is obtained by computing weighted differences between the spectral slopes in each band [30].

$$\mathrm{d_{WSS}} = \frac{1}{M}\sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K_B-1} W(j,m)\left|S_x(j,m)-\widehat{S_x}(j,m)\right|}{W(j,m)}, \tag{26}$$

where S_x and $\widehat{S_x}$ denote the spectral slope of the clean and enhanced signal, respectively. $W(j,m)$ represents the weight for the j-th band and the m-th frame. It penalizes heavily differences in spectral peak locations, i.e., formants.

For fwSNRseg, a higher score indicates better speech quality with less speech distortion. For $\mathrm{d_{WSS}}$, a lower value represents a better result. In Tables II and III, we present the upper bound of fwSNRseg and the lower bound of $\mathrm{d_{WSS}}$ for the ideal case, which is obtained by substituting the *local* SNR for the *a priori* SNR estimator. The results confirm that the proposed approach still preserves the speech characteristic even in low SNR conditions. Moreover, it outperforms the conventional DD approach for all noise types and SNR conditions. As shown in Tables II and III, the proposed *Case 2*, *Case 3*, and *AoU* approaches show good performance since the enhanced BC-based estimator adapts the *a priori* SNR estimator properly by its weighting factors. In addition, the proposed *AoU* approach is the one closest to the upper-bound of fwSNRseg and the lower bound of $\mathrm{d_{WSS}}$.

Comparing to results given in *Case 2* and *Case 3*, a temporal smoothing reduces $\mathrm{d_{WSS}}$ but it also reduces fwSNRseg. However, the AoU successfully controls the contribution of AC and BC signals with respect to local SNRs. In both the mixed noise and practical automobile noise conditions, the proposed *AoU* approach show the best speech quality with less speech distortion perceptually.

2) Subjective Measures: The speech quality was evaluated by performing the mean opinion score (MOS) test. The scores are in the range of 1 (*bad*) to 5 (*excellent perceived quality of speech*). We randomized the conditions and the order of test signals for each trial. Twelve participants (2 female and 10 male) gave scores to each enhanced speech and the averaged values are shown in Table IV.

The test results proved the perceived quality improvement of the proposed method comparing with the other approaches in most cases. Although all systems showed good and comparable quality in high SNR conditions, the proposed method showed the best performance in low SNR conditions. The subjective listening test confirmed the superiority of the proposed algorithm.

VII. CONCLUDING REMARKS

We proposed an efficient *a priori* SNR estimator by integrating an AC and a BC microphone. With the help of equalized BC speech, the feasibility of the proposed algorithm can be extended to severe low-SNR non-stationary noisy environments. The noise signal was efficiently suppressed by the proposed *a priori* SNR estimator while minimizing speech distortion. The time-frequency varying weighting factors resulting from an MMSE criterion improved the capability of eliminating the residual noise. The proposed approach also quickly estimated *a priori* SNRs even in transition regions. The weighting factors was combined by adopting a utility measure criterion. The quality of enhanced speech was improved especially in low input SNR conditions.

APPENDIX A
ACCURACY ANALYSIS OF THE ML ESTIMATE

It is interesting to see the estimation accuracy of the ML estimate $\xi_a^{(2)}(\ell,k)$ using the additive noise model if there is no assumption on the correlation between speech and noise components. $|Y_a(\ell,k)|^2$ in the additive noise model is expressed as $|S_a(\ell,k)|^2+|N_a(\ell,k)|^2+2|S_a(\ell,k)||N_a(\ell,k)|\cos\phi(l,k)$,

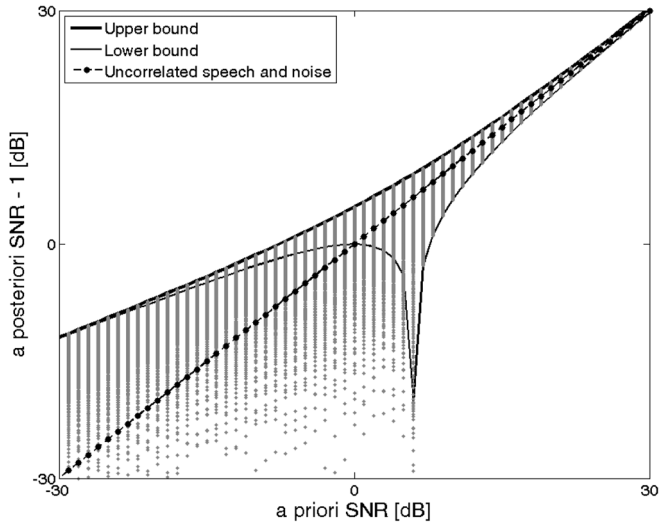

Fig. 7. Distribution of $\xi_a^{(2)}(\ell,k) = \gamma_a(\ell,k) - 1$ as a function of *a priori* SNR $\xi_a(\ell,k)$. Bounds for three cases of the phase difference ϕ. (The gray dots represent the distribution when ϕ in Eq. (27) varies from $-\pi$ to π.).

where ϕ denotes the phase difference between clean and noise signal [15]. Thus, the relationship between $\xi_a(\ell,k)$ and $\xi_a^{(2)}(\ell,k)$ can be expressed as

$$\xi_a^{(2)}(\ell,k) = \xi_a(\ell,k) + 2\sqrt{\xi_a(\ell,k)}\cos\phi(\ell,k). \qquad (27)$$

The gray dots in Fig. 7 represent the distribution when the phase difference between clean and noise signal, ϕ, in Eq. (27) varies from $-\pi$ to π.

Even the case of knowing the true clean speech and true noise signal, the estimation error still exists because of the phase difference. The upper bound of the difference occurs if $\phi(\ell,k) = 0$ (*bold solid line*) and the lower bound if $\phi(\ell,k) = \pi$ (*solid line*). In the case of $\phi(\ell,k) = \pi/2$ (*dashed line with point*) where speech and noise are uncorrelated, $\xi_a^{(2)}(\ell,k)$ becomes a perfect estimate. Therefore, the second term in Eq. (27) represents the correlation between speech and noise signals. Consequently, the uncertainty or estimation accuracy of $\xi_a^{(2)}(\ell,k)$ is related to the amount of correlation between speech and noise components. Although $\xi_a^{(2)}(\ell,k)$ is sufficiently reliable in high SNR conditions, its credibility decreases in low SNR conditions.

ACKNOWLEDGMENT

Thanks to Huajun Yu (yu@ifn.ing.tu-bs.de) and Balázs Fodor (fodor@ifn.ing.tu-bs.de) of the Institute for Communications Technology at Technische Universität Braunschweig, Braunschweig, Germany, to help recording the database in an automobile environment.

REFERENCES

[1] "Public safety communications technical report: Speech intelligibility and detection of voice characteristics," Dept. of Homeland Security, Washington, DC, USA, Tech. Rep. DHS-TR-PSC-08-05, Aug. 2008.

[2] D. J. Atkinson, S. D. Voran, and A. A. Catellier, "Intelligibility of the adaptive multi-rate speech coder in emergency-response environments," U.S. Dept of Commerce. Nat. Telecomm. and Inf. Admin., USA, Tech. Rep. NTIA Report, Mar. 2013, pp. 13–493.

[3] T. Dekens and W. Verhelst, "Body conducted speech enhancement by equalization and signal fusion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2481–2492, Dec. 2013.

[4] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, "Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling," *J. Speech Commun.*, vol. 50, pp. 228–243, Mar. 2008.

[5] P. Kechichian and S. Srinivasan, "Model-based speech enhancement using a bone-conducted signal," *J. Acoust. Soc. Amer. (JASA) Exp. Lett.*, vol. 131, pp. EL262–EL267, Feb. 2012.

[6] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in *Proc. Midwest Sympo. Circuits Syst. (MWSCAS)*, Covington, KY, USA, Aug. 2005, vol. 2, pp. 1847–1850.

[7] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Vancouver, BC, Canada, Aug. 2006, pp. 426–431.

[8] H. S. Shin and H.-G. Kang, "Bone-conduction speech enhancement using a speaker-independent filter," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Kota Kinabalu, Malaysia, Jan. 2014, pp. 327–328.

[9] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement," in *Proc. IEEE Workshop Autom. Speech Recogn. Understand. (ASRU)*, St. Thomas, U.S. Virgin Islands, Nov. 2003, pp. 249–254.

[10] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, and A. Acero, "A graphical model for multi-sensory speech processing in air-and-bone conductive microphones," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Lisbon, Portugal, Sep. 2005, pp. 2361–2364.

[11] M. Zhu, H. Ji, F. Luo, and W. Chen, "A robust speech enhancement scheme on the basis of bone-conductive microphones," in *Proc. Int. Workshop Signal Design and Its Applications in Commun. (IWSDA)*, Chengdu, China, Sep. 2007, pp. 353–355.

[12] S. Srinivasan and P. Kechichian, "Utility of auxiliary sensor data for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7294–7298.

[13] H. Yu and T. Fingscheidt, "A data-driven post-filter design based on spatially and temporally smoothed a priori SNR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 137–140.

[14] F. Chen and Philipos C. Loizou, "Impact of SNR and gain-function over- and under-estimation on speech intelligibility," *Speech Commun.*, vol. 54, pp. 272–281, Sep. 2012.

[15] C. Plapous, C. Marro, and P. Scalart, "Reliable a posteriori signal signal-to-noise ratio features selection," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2005, pp. 66–69.

[16] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement," in *IEE Proc. - Vision, Image, Signal Process.*, Jun. 2000, vol. 147, pp. 247–253.

[17] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in *Proc. ITG Speech Commun. Symp. (ITG Fachtagung Sprachkommunikation)*, Braunschweig, Germany, Sep. 2012, pp. 26–28.

[18] 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, Dec. 1984.

[19] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[20] M. S. Rahman, A. Saha, and T. Shimamura, "Low-frequency band noise suppression using bone conducted speech," in *Proc. IEEE Pacific Rim Conf. Commun., Comput., Signal Process. (PacRim)*, Victoria, BC, Canada, Aug. 2011, pp. 520–525.

[21] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct filtering for air- and bone-conductive microphones," in *Proc. IEEE Workshop Multimedia Signal Process. (MMSP)*, Siena, Italy, Sep. 2004, pp. 363–366.

[22] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[23] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.
[24] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
[25] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Lett.*, vol. 11, no. 4, pp. 450–453, Apr. 2004.
[26] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 504–512, Jul. 2001.
[27] T. Lan, D. Erdogmus, U. Ozertem, and Y. Huang, "Estimating mutual information using gaussian mixure model for feature ranking and selection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Montreal, QC, Canada, Jul. 2005, pp. 5034–5039.
[28] "Noisex-92 database," [Online]. Available: http://spib.rice.edu/spib/select_noise.html
[29] Objective measurement of active speech level, "ITU-T Recommendation," p. 56, 1993.
[30] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, Taylor & Francis, 2007.

Ho Seon Shin received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2008, 2010, and 2015, respectively. She served her internship at the Institute for Communications Technology at Technische Universität Braunschweig, Germany, from March to September of 2012. Her research interests include speech signal processing, speech enhancement, and speech reinforcement.

Tim Fingscheidt (S'93–M'98–SM'04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from RWTH Aachen University, Germany. He further pursued his work on joint speech and channel coding as a consultant in the Speech Processing Software and Technology Research Department at AT&T Labs, Florham Park, NJ, USA. In 1999, he entered the Signal Processing Department of Siemens AG (COM Mobile Devices) in Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology in Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he is Full Professor at the Institute for Communications Technology at Technische Universität Braunschweig, Germany. His research interests are speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures. Dr. Fingscheidt received several awards, among them a prize of the Vodafone Mobile Communications Foundation in 1999, and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG). From 2008 to 2010 he served as Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and from 2011 until 2015 as member of the IEEE Speech and Language Processing Technical Committee.

Hong-Goo Kang (M'02) received the B.S., M.S., and Ph.D. degrees from Yonsei University, Seoul, Korea, in 1989, 1991, and 1995, respectively. From 1996 to 2002, he was a Senior Technical Staff Member at AT&T Labs-Research, Florham Park, NJ. In 2002, he joined the Department of Electrical and Electronic Engineering, Yonsei University, where he is currently a Professor. He actively participated in international collaboration activities for making new speech/audio coding algorithms standardized by ITU-T and MPEG. His research interests include speech/audio signal processing, array signal processing, pattern recognition, and human computer interface. He was a vice chair of technical program committee in INTERSPEECH2004 held in Jeju island, Korea. He is a technical reviewing committee member of the ICASSP and INTERSPEECH conferences.