Bidirectional OM-LSA Speech Estimator for Noise Robust Speech Recognition

Yasunari Obuchi, Ryu Takeda, and Masahito Togami

Central Research Laboratory, Hitachi Ltd.
1-280 Higashi-koigakubo, Kokubunji, Tokyo, Japan
yasunari.obuchi.jx@hitachi.com

***Abstract*—A new speech enhancement method using bidirectional speech estimator is introduced. A widely-known speech enhancement method using the optimally-modified log spectral amplitude (OM-LSA) speech estimator is re-modified under the assumption that the frame-synchronous estimation is not essential in some of the speech recognition applications. The new method utilizes two separate flows of the speech gain estimation, one is along the forward direction of time and the other along the backward direction. A simple look-ahead estimation mechanism is also implemented in each flow. By taking the average of these two gains, the speech estimation becomes more robust under various noise conditions. Evaluation experiments using the artificial and real noisy speech data confirm that the speech recognition accuracy can be greatly improved by the proposed method.**

I. Introduction

Speech recognition under adverse conditions has been one of the hottest topics in the speech community. Numerous efforts have been made to improve the speech recognition accuracy, and various methods were proposed. In some cases, a modified feature extraction algorithm (e.g. RASTA [1]) was introduced or standard MFCC features were normalized before decoded (e.g. cepstral mean normalization (CMN)). In other cases, the acoustic model was adapted to the noisy environment using the noise (e.g. parallel model combination (PMC [2])) or noisy speech (e.g. maximum likelihood linear regression (MLLR [3])) data. It was also proposed to modify the decoder itself to deal with the uncertainty of the noise reducing module [4].

The problem of the above-mentioned methods is that the speech recognition system itself must be modified to adopt the new algorithm. If a new type of feature set was introduced, the whole acoustic model must be re-trained. Obviously, model adaptation is strongly tied with the recognition system. However, state-of-the-art speech recognition systems for the commercial use are made of very complex subsystems, and it is difficult to modify the whole system without reducing its stability. In contrast, the development of various speech recognition applications would be accelerated if the robustness under noisy conditions can be achieved without touching the recognition system; even the use of cloud speech recognition services would be a feasible option. Although much efforts have been made to standardize the feature for distributed speech recognition [5], the best and simplest way is to reconstruct the speech waveform after reducing the influence of noise.

Speech enhancement from noise has also been an active research area. One of the most successful achievements was the optimally-modified log spectral amplitude (OM-LSA) speech estimator with the minima controlled recursive averaging (MCRA) noise estimator, developed by Cohen and Berdugo [6]. The effectiveness of OM-LSA in terms of high noise suppression and low speech distortion has been confirmed by the objective and subjective tests [7], so it was a good starting point of our research.

When we apply OM-LSA to speech recognition, there is too strict a constraint in it. OM-LSA is a frame-synchronous algorithm, meaning that no future information is used to process the current frame data. Such a constraint is important if it is applied to real-time speech communication. However, there are many speech recognition applications in which a few second latency is allowed. In such cases, using future information makes the speech estimation more robust. In particular, robustness can be achieved near the sudden change of speech signal, which cannot be distinguished from the sudden noise unless observing the subsequent signals. In addition, the OM-LSA estimation of the current frame always relies on the estimation of the previous frames, hence the error propagates along time accumulatively if the estimation fails at a certain point.

In this paper, we propose a new speech estimator for noise robust speech recognition, which takes advantage of using the future information. A look-ahead estimator provides stability of the estimation near the sudden change of speech, and time-inverse propagation of the estimated information reduces the error accumulation. The proposed estimator uses both forward and backward recursive estimation, and referred to as the bidirectional OM-LSA speech estimator.

In the remainder of this paper, we first review the algorithm of the OM-LSA estimator in the next section. The details of the bidirectional OM-LSA estimator will be described in section III. Experimental results will be shown in Section IV, and the conclusions are given in the final section.

II. Overview of OM-LSA

Suppose that the observed signal is the mixture of the clean speech and noise. Assuming that both signals are not correlated with each other, the majority of the previous work did not take the phase information into account and were realized

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

in the spectral domain. Early work in this domain includes minimum mean square error short-term spectral amplitude (MMSE-STSA) estimator proposed by Ephraim and Malah [8], which estimates the gain of each time-frequency bin by the criterion of minimizing the mean-square error of STSA. Later, they modified the original MMSE-STSA estimator by replacing the STSA error minimization with the log-spectral amplitude (LSA) error minimization [9]. Following that, Cohen [6] introduced OM-LSA, which introduces the geometric mean of the gains associated with the speech presence and absence.

OM-LSA starts with the standard procedure in which the sampled signals are divided into overlapping frames, discontinuity at the frame edge is smoothed by applying a Hanning window, and a fast Fourier transform (FFT) is executed to obtain the signals in the frequency domain. In the frequency domain, speech enhancement is formalized as an estimation process of the gain as

$$|\hat{X}(k,l)|^2 = G(k,l)|Y(k,l)|^2 \tag{1}$$

where (k,l) are the frequency and frame indices, $Y(k,l)$ is the observed signal, $G(k,l)$ is the gain to be estimated, and $X(k,l)$ is the estimated signal. In the OM-LSA framework, the gain $G(k,l)$ is defined as the geometric mean of the two gains associated with speech presence and absence as follows.

$$G(k,l) = [G_H(k,l)]^{p(k,l)} G_{min}^{1-p(k,l)} \tag{2}$$

where $G_H(k,l)$, the gain associated with speech presence, is estimated by the LSA estimator, and G_{min}, the gain associated with speech absence, is a pre-defined constant. $p(k,l)$ is the speech presence probability.

Once the power spectrum $|\hat{X}(k,l)|^2$ is estimated, the phase information of the original signal is added, and then the frame signal in the time domain is recovered by inverse FFT and overlap-adding. The flow of the standard OM-LSA speech estimator is illustrated by the blocks connected by the dotted arrows in fig. 1.

III. Bidirectional OM-LSA with look-ahead estimation

The basic idea of our method is that we introduce two parallel speech estimators, as shown in fig. 1. One estimator utilizes the forward recursive estimation and the other utilizes the backward recursive estimation. Even if one of the estimators outputs erroneous signal, it can be compensated by the output of the other estimator.

Once the output of the FFT module was duplicated, each of them is sent to the noise estimation module. Noise estimation based on the MCRA algorithm [6] is formulated as

$$\begin{aligned}
\sigma_N^2(k,l) &= \alpha \sigma_m^2(k,l) && (3)\\
\sigma_m^2(k,l) &= \tilde{c}_1 \sigma_m^2(k,l-1) + (1-\tilde{c}_1)|Y(k,l)|^2 && (4)\\
\tilde{c}_1 &= c_1 + (1-c_1)q(k,l) && (5)
\end{aligned}$$

where α is the noise suppression coefficient that controls the balance between noise reduction and speech distortion, $q(k,l)$

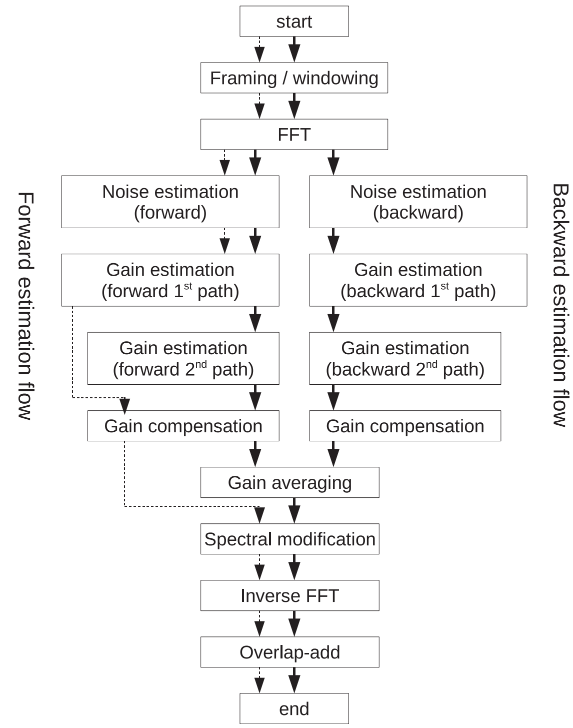

Fig. 1. Schematic diagram of bidirectional OM-LSA speech estimator with look-ahead estimation. The blocks connected by the dotted arrows construct the flow of standard OM-LSA.

is the speech presence probability calculated by the minimum tracking, and c_1 is a pre-defined weight parameter. The inter-frame dependency appears in eq. (4), where the $l-1$-th frame is used to estimate l-th frame since it is forward estimation. In backward estimation, $l-1$ must be replaced by $l+1$.

Next, the first path of gain estimation is executed. It is the same as standard LSA estimation defined by

$$\begin{aligned}
G_{f1}(k,l) &= f(\xi_{f1}(k,l), \gamma(k,l)) && (6)\\
f(\xi,\gamma) &= \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\gamma\xi/(1+\xi)}^{\infty}\frac{e^{-1}}{t}dt\right) && (7)\\
\gamma(k,l) &= \frac{|Y(k,l)|^2}{\sigma_N^2(k,l)}, && (8)
\end{aligned}$$

where ξ_{f1} is the a priori SNR and γ is the a posteriori SNR. The subscripts $f1$ stand for the first path of forward estimation. The a priori SNR is estimated by

$$\xi_{f1}(k,l) = c_2 G_{f1}^2(k,l–1)\gamma(k,l–1) + (1–c_2)max\{\gamma(k,l)–1, 0\}, \tag{9}$$

where c_2 is a pre-defined parameter. It should be noted again that $l-1$ must be replaced by $l+1$ in the backward estimation flow as follows.

$$G_{b1}(k,l) = f(\xi_{b1}(k,l), \gamma(k,l)) \tag{10}$$

$$\xi_{b1}(k,l) = c_2 G_{b1}^2(k,l–1)\gamma(k,l–1) + (1–c_2)max\{\gamma(k,l)–1, 0\}. \tag{11}$$

where the subscripts $b1$ denote the first path of backward estimation.

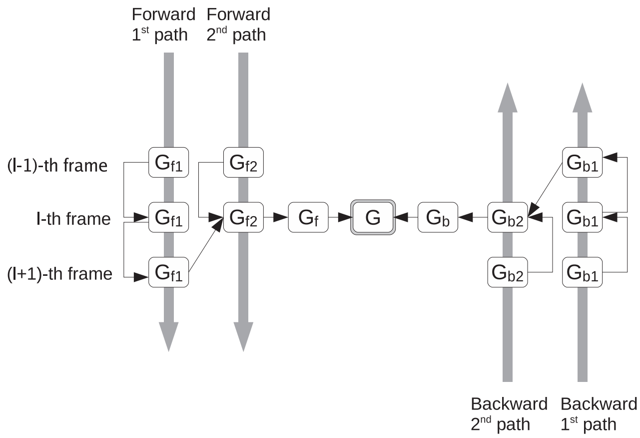

Fig. 2. Relationship of the paths in the forward and backward flows.

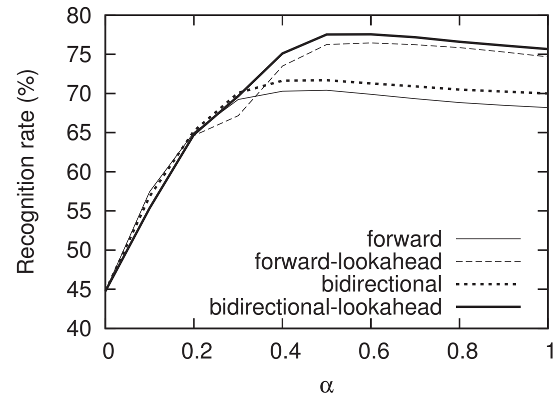

Fig. 3. Experimantal results for the artificial data obtained by standard OM-LSA (forward) and three improved methods. Recognition rates are the average under four SNR conditions.

In the case of standard OM-LSA, $G_{f1}(k,l)$ is directly put into eq. (2), where it is substituted for $G_H(k,l)$, and

$$p(k,l) = [1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi_{f1}(k,l))e^{-\nu(k,l)}]^{-1} \qquad (12)$$

$$\nu(k,l) = \gamma(k,l)\xi_{f1}(k,l)/(1+\xi_{f1}(k,l)) \qquad (13)$$

where $q(k,l)$ is the a priori probability for speech absence.

In contrast, our method introduces the second path of gain estimation, where the information of the next frame is used to estimate the current frame. The second path, referred to as *look-ahead* estimation, is represented by the subscripts $f2$ and defined as follows.

$$G_{f2}(k,l) = f(\xi_{f2}(k,l), \gamma(k,l)) \qquad (14)$$

$$\xi_{f2}(k,l) = c_2 G_{ave}\gamma_{ave} + (1-c_2)\max\{\gamma_{ave}(k,l)-1, 0\} \qquad (15)$$

$$G_{ave} = \{G_{f2}(k,l-1) + G_{f1}(k,l+1)\}/2 \qquad (16)$$

$$\gamma_{ave} = \{\gamma(k,l-1) + \gamma(k,l+1)\}/2. \qquad (17)$$

Although the whole bidirectional system requires all data to be observed and stored before the estimation begins, it is possible to use look-ahead estimation alone, by which only one frame latency is required. Similar idea was proposed by Cohen [10], in which the estimated a priori SNR is partly replaced by the non-recursive average from the near past to the near future. In our method, which uses fewer adjustable parameters than Cohen's, the future gain is estimated recursively in the first path, and fed back to the second path.

The second path is also applied to the backward flow, where the subscripts are changed to $b2$. The output of the second path, G_{f2} and G_{b2}, are then put into eq. (2),

$$G_f(k,l) = [G_{f2}(k,l)]^{p_{f2}(k,l)} G_{min}^{1-p_{f2}(k,l)} \qquad (18)$$

$$G_b(k,l) = [G_{b2}(k,l)]^{p_{b2}(k,l)} G_{min}^{1-p_{b2}(k,l)} \qquad (19)$$

where $p_{f2}(k,l)$ and $p_{b2}(k,l)$ are obtained by eq. (12) using ξ_{f2} or ξ_{b2} instead of ξ_{f1} respectively. Finally,

$$G(k,l) = \{G_f(k,l) + G_b(k,l)\}/2. \qquad (20)$$

is the gain of the proposed method, and the standard procedure to reconstruct the waveform follows. The relationship of these paths in the two flows is illustrated in fig. 2.

IV. Experimental Results

The proposed method was evaluated by two sets of speech recognition experiments. The first set consists of the artificial noisy speech data. The noise was added with various SNRs to the clean speech, so we can compare the data and results with various SNRs. The second set consists of the well-known public database, CENSREC-2. Since all of the data in CENSREC-2 were recorded in real noisy environments, we can obtain more reliable evaluation results, although we cannot analyze them using the clean speech. It is also helpful that the results can be easily compared with the literature.

Pre-defined parameters such as G_{min}, $q(k,l) = q_0$, c_1, and c_2 were empirically determined, but they are fixed throughout all the experiments described below. The window size and window shift of speech enhancement were 62 msec and 32 msec respectively, which are not the same as the speech recognition experiments.

A. Experiments using artificial data

The evaluation data of the first set of experiments were made with our proprietary database. The clean dataset consists of 4,000 isolated words uttered by 10 male and 30 female speakers. Each word is a Japanese family name such as "suzuki" and "sato." The noise was recorded in a computer room, and added to the clean dataset with the SNR of 15dB, 10dB, 5dB, and 0dB. Training data for the acoustic model were made of phonetically balanced sentences and words uttered by 200 male and 282 female speakers in a quiet room. The total length of the training data is approximately 240 hours. All data were recorded with 16kHz sampling frequency.

The acoustic model is made of tied-state left-to-right triphone HMMs. Each of 2,563 states has 16 Gaussian mixtures. To reduce the execution time, all of 41,008 mixtures are quantized to 1,024 codewords in nine subspaces separately using

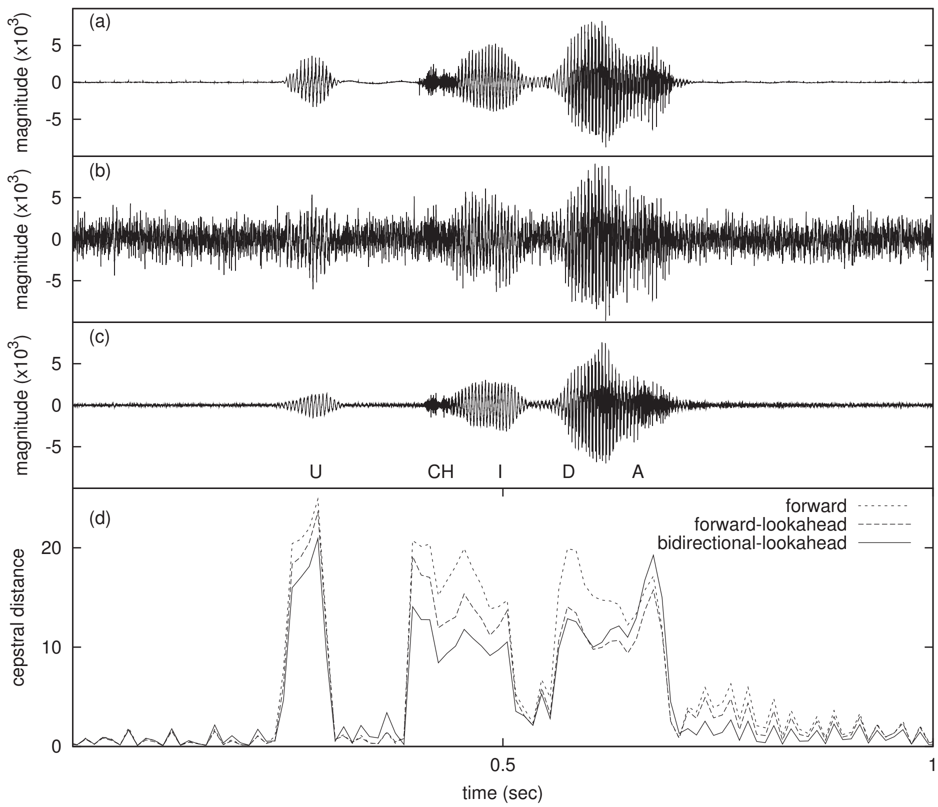

Fig. 4. Example of speech estimation. The speaker said "UCHIDA." (a) Clean speech (b) Noisy speech (0dB SNR) (c) Estimated speech by the proposed method (d) Cepstral distances between MFCC feature vectors of clean and estimated speeches. The speech estimated by "forward" and "forward-lookahead" were recognized as "ISHIDA," due to the large mismatch around "CH." The speech estimated by "bidirectional-lookahead" was recognized correctly.

subvector quantization. The recognition task is the isolated word recognition of 100 Japanese family names. Each entry of the vocabulary is made of three to eight phonemes (average 6.1 phonemes). The feature vector consists of 13 MFCCs including 0-th MFCC, with their first and second order time derivatives. Cepstral mean normalization was applied to the training and test data, where the cepstral mean was calculated for each utterance. The recognition rate for the clean data was 99.2%

Figure 3 shows the average recognition rates of the four SNR conditions obtained by standard OM-LSA (thin solid line denoted by "forward" in the figure) and three improved methods, as the function of the noise suppression coefficient α. The dashed line denoted by "forward-lookahead" represents the experiments using $G_f(k,l)$ instead of $G(k,l)$ in eq. (1). The dotted line denoted by "bidirectional" represents the experiments using eq. (20) but omitting the second path of eqs. (14) – (17). The thick solid line denoted by "bidirectional-lookahad" represents the experiments with all improvements.

In fig. 3, the baseline recognition rate of 44.8% corresponds to the point at $\alpha = 0$. The recognition rate ascends rapidly as α increases in any case, and takes the peak at $\alpha = 0.5$ or $\alpha = 0.6$. In most cases, all three improved methods provide higher recognition rates than standard OM-LSA. In particular, introducing look-ahead estimation brings significant improvement.

TABLE I
DETAILED COMPARISON OF THE RECOGNITION RATES FOR THE ARTIFICIAL DATA.

SNR (dB)	baseline	forward		bidirectional	
		normal	look-ahead	normal	look-ahead
15	89.6	95.7	96.2	96.0	96.6
10	63.9	87.2	90.4	88.0	91.0
5	21.0	66.4	75.0	69.1	77.1
0	4.5	32.4	44.1	33.6	45.5
overall (rel.imp.)	44.8	70.4 (46.4)	76.5 (57.4)	71.7 (48.7)	77.6 (59.4)

Table I shows the details of the recognition rates with various SNRs. The noise suppression coefficient α was fixed at the optimal point for each method. It can be seen that the recognition rates under low SNR conditions were greatly improved. In this table, the relative improvement was defined as

$$\text{rel. imp.} = \frac{r - r_0}{100.0 - r_0} * 100.0 \tag{21}$$

where r and r_0 are the recognition rates of the modified and baseline experiments, respectively. The largest relative

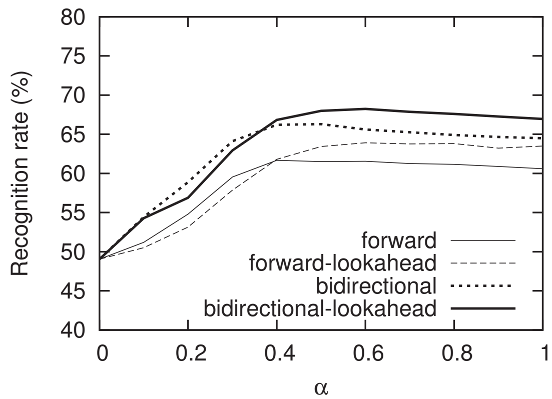

Fig. 5. Experimantal results for CENSREC-2 obtained by standard OM-LSA (forward) and three improved methods without CMN. Recognition rates are the average under seven conditions.

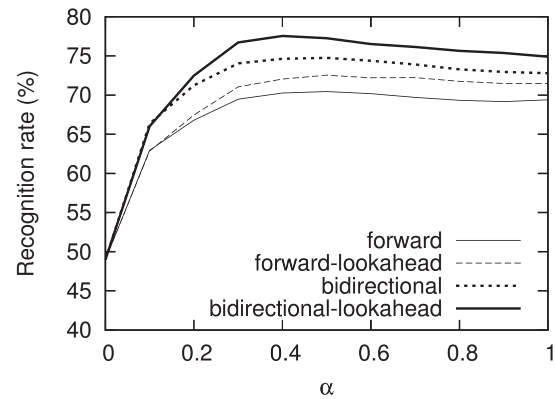

Fig. 6. Experimantal results for CENSREC-2 obtained by standard OM-LSA (forward) and three improved methods with CMN. Recognition rates are the average under seven conditions.

improvement of 59.4% was obtained by the "bidirectional-lookahead" method, which is 13.0 point higher than standard OM-LSA and 2.0 points higher than OM-LSA with look-ahead estimation only.

To illustrate how the speech signal is estimated from the noisy speech input, an example utterance and the detailed analysis of its estimation process are shown in fig. 4. In this example, the original speech (a), an utterance of "UCHIDA," was mixed with the noise at the SNR of 0dB. The noisy speech (b) was modified to (c) by the proposed "bidirectional-lookahead" method. At the bottom of the figure, cepstral distances between the clean and estimated speeches are plotted. The cepstral distance of the l-th frame was defined by

$$D_l = \sum_{i=0}^{12} (x(i,l) - \hat{x}(i,l))^2 \qquad (22)$$

where $x(i,l)$ is the i-th MFCC of the clean speech and $\hat{x}(i,l)$ is the i-th MFCC of the estimated speech. Both clean and estimated MFCCs had been normalized by CMN beforehand.

It can be seen in fig. 4 that there are large mismatches at the positions of all phonemes, and the mismatch near "A" was not reduced by any method. The mismatch near "D" was reduced greatly by the look-ahead estimation, and no further improvement was obtained by the bidirectional estimation. However, it did not help the recognition very much because there was no confusing word that differs only at this phoneme from the uttered word. The mismatch around "CHI" was gradually reduced by the look-ahead and then bidirectional estimation, and the correct recognition was achieved only by the bidirectional-lookahead estimation. It is because the vocabulary includes a confusing word "ISHIDA." The clues to distinguish "UCHIDA" and "ISHIDA" exist at "U/I" and "CH/SH," but the mismatch at the former position was not reduced by any method, and hence the latter is the only chance to avoid the misrecognition.

B. Experiments using CENSREC-2

The second set of experiments was carried out using CENSREC-2 [11]. CENSREC-2 is a combination of a speech database and an evaluation framework. The database consists of the training and test data, and the evaluation framework consists of various scripts for the experiments with HTK [12]. The voices were recorded when the car was and was not running, by the close-talk and hands-free microphones at the same time. However, we used the close-talk data that were recorded when the car was not running for HMM training, and the hands-free data that were recorded when the car was running for evaluation. This combination was defined as **Condition 4** by the creator of CENSREC-2. The test data consist of 2,058 utterances uttered by 19 male and 12 female speakers, each of which is made of one to seven connected digits pronounced in Japanese. The training data consists of 2,737 utterances uttered by 33 male and 40 female speakers. All data were recorded with 16kHz sampling frequency

The acoustic model was trained automatically using the CENSREC-2 scripts. The feature vector consists of 12 MFCCs and log-energy with their first and second order time derivatives. The experiments using this model belong to **Category 0** defined in the attached document of CENSREC-2. We also made another model using CMN. The experiment using this model belongs to **Category 1**, but the change of the script was only one line, so it is very close to **Category 0**.

Figure 5 shows the average recognition rates of CENSREC-2 experiments without CMN. Corresponding results with CMN were plotted in fig. 6. The four methods evaluated in these experiments are the same as in fig. 3. It can be seen again that bidirectional and look-ahead estimation improved the recognition rate. However, contrary to the previous experiments, the dotted line representing "bidirectional" lies above the dashed line representing "forward-lookahead." It means that introducing bidirectional estimation was more effective than look-ahead estimation under the running car condition. Although the baseline recognition rates with and without

TABLE II
DETAILED COMPARISON OF THE RECOGNITION RATES FOR CENSREC-2.

speed	condition	baseline	CMN	Li [13]	Proposed
low	normal	63.8	71.5	86.2	89.3
	fan on	56.8	54.1	83.3	90.7
	audio on	51.4	56.3	69.2	61.5
	window open	46.6	44.8	70.8	79.2
high	normal	43.3	42.6	73.0	81.4
	fan on	40.0	33.7	70.3	79.6
	audio on	41.1	39.2	62.4	61.1
overall (relative improvement)		49.1	49.0 (-0.2)	73.7 (48.3)	77.6 (56.0)

CMN are very close to each other, greater improvement was obtained with CMN regardless of the estimation method. The recognition rate with CMN and bidirectional-lookahead estimation was 77.6% when $\alpha = 0.4$, which means 56.0% relative improvement from the baseline.

Since CENSREC-2 is a publicly available database, we also compared our results with literature. In [13], it was said that a non-linear spectral contrast stretching method outperforms some of the well known methods such as LSA [9] and ETSI advanced front-end [5], so the performance of their method could be a good benchmark.

Table II shows the detailed comparison between the proposed method and Li's non-linear spectral contrast stretching. There are seven subsets representing various car speeds and in-car conditions, and the proposed method outperformed Li's method in five of them. The relative improvement of the proposed method (56.0%) was 7.7 points higher than Li's (48.3%). The two subsets in which the proposed method was worse than Li's were both under "audio on" condition. We can infer that the proposed method is less effective for the noises including music and human voices.

V. CONCLUSION

In this paper, we introduced a new speech enhancement method, which is based on the OM-LSA speech estimator, and optimized for pre-processing of speech recognition. In the proposed method, the gain estimation in the spectral domain is separated into two flows: forward recursive estimation and backward recursive estimation. Even if the estimation error accumulates in one of two flows, it can be compensated by the estimation of the other flow. In each flow, additional improvement was obtained by introducing a look-ahead estimation that is expected to stabilize the gain estimation. The two gains obtained by the forward and backward estimation flows are averaged and multiplied to the input signal, resulting in the reconstruction of the estimated speech signal, which is fed into the speech recognizer.

The effectiveness of the proposed method was confirmed by two sets of evaluation experiments. The experiments using artificial data, which was made by combining the clean speech and computer room noise, revealed that the look-ahead estimation was particularly effective under such conditions, whereas an additional improvement can be obtained by the bidirectional estimation. An opposite trend was observed in the experiments using CENSREC-2, a publicly available evaluation environment. The improvement by the bidirectional estimation was larger than that of the look-ahead estimation, but we had the consistent result that the proposed speech estimator equipped with these two improvements was the best among various methods. In the experiments using CENSREC-2, it was also confirmed that the proposed method outperformed other noise robust speech recognition methods found in the literature. A more detailed analysis showed that the proposed method could not outperform the existing method only under "audio-on" conditions. It would be an important future work to investigate the mechanism in which the proposed method fails to suppress the audio noise efficiently, and provide a solution for such cases.

REFERENCES

[1] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis," *ICSI Technical Report* TR-91-069, UC Berkeley, 1991.

[2] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language* Vol.9, pp.289-307, 1995.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol.9, No.2, pp.171-185, 1995.

[4] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Denver, FL, USA, 2002.

[5] Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 050, V1.1.5, 2007.

[6] I. Cohen and B. Berdugo, "Speech enhancement for Non-stationary Noise Environments," *Signal Processing*, Vol.81, pp.2403-2418, 2001.

[7] S. Gannot, "Speech enhancement: application of the Kalman filter in the estimate-maximize (EM) framework," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Ed., Berlin: Springer, 2005, pp.161-198.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.32, No.6, pp.1109-1121, 1984.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.33, No.2, pp.443-445, 1985.

[10] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Processing Letters*, Vol.11, No.9, pp.725-728, 2004.

[11] S. Nakamura, M. Fujimoto, and K. Takeda, "CENSREC2: Corpus and evaluation environments for in car continuous digit speech recognition," *Proc. Interspeech2006*, Pittsburgh, PA, USA, 2006.

[12] HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/

[13] W. Li and H. Bourlard, "Non-linear spectral contrast stretching for in-car speech recognition," *Proc. Interspeech2007*, Antwerp, Belgium, 2007.