Noise suppression method for preprocessor of time-lag speech recognition system based on bidirectional optimally modified log spectral amplitude estimation

Yasunari Obuchi*, Ryu Takeda and Masahito Togami

Central Research Laboratory, Hitachi Ltd.,
1–280 Higashi-koigakubo, Kokubunji, 185–8601 Japan

(Received 23 May 2012, Accepted for publication 22 November 2012)

Abstract: In this paper, we propose a new noise suppression method, that is best used as a preprocessor for time-lag speech recognition. Assuming that a time lag of a few seconds is acceptable in various speech recognition applications, the proposed method is realized as a combination of forward and backward estimation flows over time. Each estimation flow is based on the optimally modified log spectral amplitude (OM-LSA) speech estimator, but a look-ahead estimation mechanism is additionally equipped to make the estimation more robust. Evaluation experiments using various databases confirm that the speech recognition accuracy can be greatly improved by adding the proposed method to the existing system.

Keywords: Noise suppression, OM-LSA, Bidirectional, Speech recognition, CENSREC-2

PACS number: 43.72.Ne [doi:10.1250/ast.34.133]

1. INTRODUCTION

Speech recognition technology has achieved great commercial success as a result of intensive research over the past few decades. State-of-the-art systems are complex and contain various components, including acoustic and language models, a feature extractor, and a graph searcher. Each of these components has been optimized through a time-consuming development process, and thus modularization is becoming more important. It is reasonable to renew one component for a specific application when new insights are provided. However, renewing the entire system is tiresome, or may be impossible if the system was provided from a third-party organization.

Although the robustness of speech recognition systems has been one of the primary research targets in recent years, many new algorithms have not yet been implemented in real applications because of the situation described above. For example, introducing a new type of feature [1,2] or applying a feature normalization algorithm [3–7] requires retraining of the acoustic model. Acoustic model adaptation [8,9] can be used only when one has access to the detailed structure of the acoustic model. Obviously, it is the most difficult approach to modify the decoder itself, as described in [10].

In this work, we tackle this problem by introducing a simple and powerful preprocessor to an existing system. The preprocessor is realized as a speech enhancement module because raw speech data is the only format which most existing speech recognition systems accept, despite efforts to standardize the feature for speech recognition [11]. As the starting point, we have chosen the optimally modified log spectral amplitude (OM-LSA) speech estimator, combined with the minima-controlled recursive averaging (MCRA) noise estimator developed by Cohen and Berdugo [12]. It was confirmed by objective and subjective tests [13] that the OM-LSA speech estimator can achieve satisfactory suppression of noise with only a slight distortion of the signal.

Our improvement starts with the fact that a preprocessor for speech recognition may be different from that for human-to-human speech communication. Some speech recognition applications can adopt a time-lag approach, in which the decoding starts after a certain amount of data has been accumulated. In such a case, the preprocessor also allows time-lag implementation, whereas a latency longer than 150 ms is unacceptable for most voice over internet protocol (VoIP) applications [14]. This is because the speech communication system is required to simply transfer the signal, whereas the speech recognition system is regarded as something that thinks for a while before responding. From this viewpoint, we improve the OM-LSA...
speech estimator with the condition that a latency as long as the utterance itself is allowed. This means that the improved algorithm can be a batch algorithm. In a batch algorithm, we can use future information (and, obviously, past information) to estimate the current signal. This would make the estimation near a sudden change in the speech signal more robust. It is also expected that the cumulative error over time will be reduced by using both forward and backward recursive processing. That is the reason why our new system is named the “bidirectional OM-LSA speech estimator” [15,16].

The remainder of the paper is organized as follows. In the next section, we briefly review the formulation of the OM-LSA speech estimator. In Sect. 3, the bidirectional OM-LSA speech estimator is described in detail. Section 4 presents various evaluation experiments performed under stationary and nonstationary noise conditions. The evaluation environments include our proprietary dataset and the CENSREC-2 [17] public evaluation framework. The last section summarizes the work and presents some future perspectives.

2. BRIEF OVERVIEW OF OM-LSA SPEECH ESTIMATOR

Statistical modeling of the OM-LSA speech estimator in the spectral domain originated from the minimum mean-square-error short-term spectral amplitude (MMSE-STSA) estimator proposed by Ephraim and Malah [18], in which the speech signal and noise are modeled as independent Gaussian random processes. As its name indicates, the estimation was based on the minimization of the mean-square error of STSA. Later, error minimization in terms of the log-spectral amplitude (LSA) was introduced [19] and then improved by introducing the explicit adjustment using the speech presence probability in each time-frequency bin. The improved version was called the OM-LSA speech estimator.

The goal of the STSA or LSA estimator is to obtain the gain function in the frequency domain:

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

where \((k,l)\) are the frequency and frame indices, \(Y(k,l)\) is the observed signal converted to the frequency domain, \(G(k,l)\) is the gain function, and \(\hat{X}(k,l)\) is the estimated signal spectrum.

After some mathematical procedures to minimize the LSA mean-square error, the gain function is obtained as follows (suffixes \(f\) and \(f1\) are added to maintain consistency with the later sections):

$$G_{f1}(k,l) = f(\xi_{f1}(k,l), \gamma_{f1}(k,l))$$

$$f(\xi, \gamma) = \frac{\xi}{1 + \xi} \exp\left(1 - \frac{1}{2} \int_{\frac{1}{1+\xi}}^{\infty} \frac{e^{-t}}{t} dt\right)$$

where \(\xi\) is the a priori SNR and \(\gamma\) is the a posteriori SNR. The a priori SNR is estimated by

$$\xi_{f1}(k,l) = c_1 G_{f1}(k,l - 1) \gamma_{f1}(k,l) + (1 - c_1) \max(\gamma_{f1}(k,l) - 1, 0)$$

where \(c_1\) is a predefined parameter. It should be noted that \(G\) and \(\gamma\) of the \((l-1)\)th frame are used here, meaning that the estimation is done recursively along the forward direction.

The a posteriori SNR is estimated with the help of noise estimation by MCRA.

$$\sigma_{\text{mm}}^2(k,l) = \alpha \sigma_{\text{mm}}^2(k,l)$$

$$\sigma_{\text{mm}}^2(k,l) = \hat{\varepsilon}_2(k,l) \sigma_{\text{mm}}^2(k,l - 1) + (1 - \hat{\varepsilon}_2(k,l))|Y(k,l)|^2$$

$$\hat{\varepsilon}_2(k,l) = c_2 + (1 - c_2) q(k,l)$$

Here, \(\alpha\) is the noise suppression coefficient that controls the balance between noise reduction and speech distortion, \(q(k,l)\) is the speech presence coefficient, \(G(k,l)\) is the gain of the OM-LSA speech estimator proposed in this paper incorporates two improvements. In this section, we describe the details of each improvement. A schematic diagram of the proposed method is illustrated in Fig. 1.

3. BIDIRECTIONAL OM-LSA SPEECH ESTIMATOR

The speech estimator proposed in this paper incorporates two improvements. In this section, we describe the details of each improvement. A schematic diagram of the proposed method is illustrated in Fig. 1.

3.1. Look-Ahead Mechanism

The simplest way to avoid erroneous estimation near a sudden change in the signal is to use future information in
addition to past information. Such an approach was already proposed by Cohen [20]; the a priori SNR is estimated partly from the average of the past and future observations. In our method, estimation is carried out in a simpler manner with fewer adjustable parameters, as depicted in the upper half of Fig. 1. The uppermost flow of the figure represents the standard procedure of the OM-LSA speech estimator described in the previous section. After obtaining $G_f$ of the $(l + 1)\text{th}$ frame, our focus moves back to the $l$th frame to execute the second path. Assuming that the second path has already been executed up to the $(l - 1)\text{th}$ frame, the gain of the target frame is estimated using the next frame of the first path and the previous frame of the second path as follows:

$$G_f(k,l) = f(\xi_f(k,l), \gamma_f(k,l))$$

(12)

$$\xi_f(k,l) = c_lG_{f,ave}(k,l)\gamma_f(k,l) + (1 - c_l)\max\{\gamma_f(k,l) - 1, 0\}$$

(13)

$$G_{f,ave}(k,l) = \{G_f(k,l-1) + G_f(k,l+1)\}/2$$

(14)

$$\gamma_f(k,l) = \{\gamma_f(k,l - 1) + \gamma_f(k,l + 1)\}/2$$

(15)

$$G_f(k,l) = [G_{f2}(k,l)]^{p_f(k,l)}G_{min}^{1-p_f(k,l)},$$

where the suffix $f2$ is used to identify the second path, $p_f$ is obtained by substituting $\xi_f$ into Eq. (10), and $G_f$ represents the final estimate. If we compare Eqs. (5) and (13), it can be confirmed that the estimation using the previous frame is replaced with the estimation using the previous and next frames while the estimation using the current frame is kept the same.

It should be noted that Eqs. (12) to (16) require no additional parameters, whereas Cohen’s method introduces some additional parameters (note that $c_l$ and $G_{min}$ were already defined by the original OM-LSA speech estimator).

The problem of parameter tuning will be discussed from the experimental viewpoint in the next section.

### 3.2. Bidirectional Estimation

The second improvement was introduced as a solution to the error accumulation problem. Since the original OM-LSA speech estimator is a recursive estimator, error might be accumulated over time. However, such large errors could be reduced by introducing the estimation along the opposite direction. This can be easily understood by considering that interpolation is usually more precise than extrapolation.

The backward estimation path, depicted in the lower half of Fig. 1, is a simple mirror image of the forward estimation path. The equations below can be derived from the equations for the forward path simply by swapping the frame indices $l + 1$ and $l - 1$.

$$G_b(k,l) = f(\xi_b(k,l), \gamma_b(k,l))$$

(17)

$$\xi_b(k,l) = c_lG_{b,ave}(k,l)\gamma_b(k,l) + (1 - c_l)\max\{\gamma_b(k,l) - 1, 0\}$$

(18)

$$G_{b,ave}(k,l) = \{G_b(k,l+1) + G_b(k,l-1)\}/2$$

(19)

$$\gamma_b(k,l) = \{\gamma_b(k,l + 1) + \gamma_b(k,l - 1)\}/2$$

(20)

$$G_b(k,l) = [G_{b2}(k,l)]^{p_b(k,l)}G_{min}^{1-p_b(k,l)},$$

Here, $p_b(k,l)$ is obtained by substituting $\xi_b$ into Eq. (10). Finally,

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

(23)

is the gain of the proposed method.
4. EVALUATION EXPERIMENTS

We carried out two sets of speech recognition experiments to evaluate the proposed method. In the first set, we use CENSREC-2, which is a combination of a database and evaluation framework. Since CENSREC-2 is publicly available, we can compare our results with those in the literature. In the second set, we use our own database and speech recognizer, in which we can control various parameters to see more details.

4.1. Experiments Using CENSREC-2

CENSREC-2 [17] comprises the training and test data and some scripts for HTK [21]. It is designed to emulate connected digit recognition in a car. The acoustic model is obtained almost automatically using the script and training data. The feature vector consists of 12 MFCCs and log-energy with their first- and second-order time derivatives. Utterance-wise cepstral mean normalization (CMN) was applied for the training and test data, hence the experiments belong to Category 1. Among the various variations, we chose the subset called Condition 4, in which the training data were recorded using a close-talk microphone when the car was not running, and the test data were recorded using a distant-talk microphone when the car was running. The training data consist of 2,737 utterances by 33 male and 40 female speakers, and the test data consist of 2,058 utterances by 19 male and 12 female speakers. All data were recorded with 16 kHz sampling frequency.

In the evaluation experiments, the standard OM-LSA speech estimator and three variations of the proposed method were applied to each test utterance, and the reconstructed waveform was used by the recognition script. The first variation uses the look-ahead mechanism only, with standard (forward) estimation. The second variation uses bidirectional estimation only, and the third variation uses both the look-ahead mechanism and bidirectional estimation.

Figure 2 shows the results of speech recognition experiments. The horizontal axis represents $\alpha$, introduced in Eq. (6). The vertical axis represents the recognition rate, which is the average over seven different conditions of the car. Although the baseline system provides a very low recognition rate (49.1%), it increases rapidly if any noise suppression method is applied. It is observed that bidirectional estimation is more effective than the look-ahead mechanism, and the best performance was obtained when we used both of them. In that case, the highest recognition rate of 77.6% was obtained when $\alpha = 0.4$. It should also be noted that the recognition rate does not change greatly for a wide range of $\alpha$.

More detailed results are shown in Table 1, in which the results by Li and Bourlard are also presented for comparison. The column “Proposed” corresponds to the bidirectional look-ahead method with $\alpha = 0.4$. In [22], it is reported that the nonlinear spectral contrast stretching method, proposed in that paper, outperforms the LSA speech estimator [19] and ETSI advanced front-end [11]. Therefore, Li and Bourlard’s method could be a good benchmark for evaluation. In Table 1, it can be seen that our method provides higher recognition rates than does Li and Bourlard’s method under five of seven conditions. The two exceptional conditions are both audio-on conditions, which suggests that our method is less effective for noise comprising human voices or music.

Figure 3 shows the results of experiments on comparing our look-ahead mechanism and that in [20]. We retained the parameter setting used in the experiments of Fig. 2, in which $\alpha$ was fixed at 0.4, and all the other parameter values were taken from [20], except that only one future frame and no adjacent frequency bin were used\(^*\). The corresponding recognition rate was 63.0%, which was even lower than that of the standard OM-LSA speech

\(^*\)We intended to use the same additional information to estimate the current time-frequency bin in both methods.
estimator. We found that the recognition rate is particularly sensitive to the nonstationarity degree parameter $\mu'$ (see [20] for the definition of $\mu'$), and carried out additional experiments using various values of $\mu'$. As Fig. 3 shows, the recognition rate became comparable to the proposed look-ahead mechanism only when the optimal value of $\mu'$ ($= 0$) was used.

4.2. Experiments Using Original Data and Speech Recognizer

Next, we carried out experiments using our original data and speech recognizer. The task is isolated word recognition, in which the target lexicon is made of 100 popular Japanese family names. The clean test data of 4,000 utterances (100 family names uttered by 10 male and 30 female speakers) were recorded in a quiet room, and either computer room noise or exhibition hall noise was added with an SNR of 15 dB, 10 dB, 5 dB, or 0 dB. The total number of test utterances was 16,000. The computer room noise was recorded by ourselves, and the exhibition hall noise was obtained by downsampling set No. 3 of the JEIDA noise database [23]. All data were prepared with 16 kHz sampling frequency.

The acoustic model was trained using 240 hours of clean speech, which comprises phonetically balanced words and sentences uttered by 200 male and 282 female speakers. The feature vector consists of 13 MFCCs, including the 0th MFCC, with their first- and second-order time derivatives. Utterance-wise CMN was applied. The acoustic model is made of tied-state left-to-right tristate triphone HMMs. Each of the 2,563 states has 16 Gaussian mixtures with diagonal covariance matrices, and subvector quantization was used to reduce the computational burden.

![Fig. 3](image3.jpg)

**Fig. 3** Recognition rates obtained with various values of the nonstationarity degree $\mu'$. The original OM-LSA speech estimator and the look-ahead method proposed in this paper do not include $\mu'$. The filled circle corresponds to the result from [20].

![Fig. 4](image4.jpg)

**Fig. 4** Recognition rates for the original data with computer room noise.

Figure 4 shows the experimental results for the data with computer room noise. The shape of the recognition rate curve is similar to that of CENSREC-2. The only difference is that the look-ahead mechanism provides greater improvement than does bidirectional estimation. However, the combination of the two methods gives the best results, as expected. The highest recognition rate is 77.6%, which is obtained with $\alpha = 0.6$.

To illustrate how the proposed method improves the speech recognition rate, an example is shown in Fig. 5, with the analysis performed in terms of the cepstral distance. The original speech, an utterance of “TAKEDA,” was contaminated with noise at an SNR of 0 dB, and then processed by either “forward,” “forward look-ahead,” or “bidirectional look-ahead” noise suppression. The framework cepstral distance between the clean and noise-suppressed speech is defined by

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

where $x(i, l)$ is the $i$th MFCC of the $l$th frame of the clean speech, and $\hat{x}(i, l)$ is the corresponding MFCC of the noise-suppressed speech. CMN was applied for clean and noise-suppressed MFCCs before calculating the cepstral distance.

From Fig. 5(d), it can be seen that the largest mismatch appears near the syllable “TA.” The large mismatch of “forward” is reduced slightly by “forward look-ahead,” and still more by “bidirectional look-ahead.” This is the reason why the utterance was misrecognized as “IKEDA” by “forward” and “forward look-ahead,” but recognized correctly as “TAKEDA” by “bidirectional look-ahead.” The mismatch of “bidirectional look-ahead” near the syllable “DA” is slightly greater than that of “forward look-ahead.” However, it does not change the recognition result because the difference is rather small and there is no conflicting word from which this utterance must be distinguished by this syllable.
Figure 6 shows the results for another set of experiments, in which exhibition hall noise was added to the speech. By listening to the noise data, we can easily note that exhibition hall noise is not as stable as computer room noise. Although the look-ahead mechanism and bidirectional estimation were introduced to reduce the degradation caused by such nonstationarity, too large changes in the noise could not be compensated perfectly. In particular, dealing with such difficult situations might cause distortion of the signal, and we carried out these experiments to investigate how the proposed method performs in such situations.

In Fig. 6, the plots of all four methods make strange shapes. As $\alpha$ increases, the recognition rate also increases at first. However, it drops rapidly as $\alpha$ becomes larger than 0.1. After $\alpha$ exceeds 0.3, it improves again but never reaches the baseline performance.

To determine what triggers the drop of the recognition rate near $\alpha = 0.3$, we selected some examples and checked the alignment given by the decoder. Figure 7 shows an example in which the noisy and noise-suppressed speeches of “yamamoto” were both misrecognized. In the case of the noisy speech, the silence part was correctly aligned, but there are some mismatches near the second “m.” In the case of the noise-suppressed speech, the first phoneme “y” was incorrectly aligned with the silence period before the utterance, whereas the remaining part was aligned correctly. This suggests that the noise suppression algorithm could not remove the time-varying noise sufficiently, and the residual musical noise somehow resembles the speech rather than the known background noise.

4.3. Silence Model Augmentation

Although our original goal was to provide a noise suppression tool for any speech recognizer, we became interested in what would happen if the speech recognizer were more robust against such mismatches. This is
equivalent to investigating how the proposed method can be effectively improved if it works in a more conservative manner in the nonspeech period. Although such improvement requires the proposed method to be combined with state-of-the-art voice activity detection, which requires some additional research, its effect can be estimated easily in experiments using the modified recognizer. Accordingly, we modified the silence model of our HMM. We simply modified the variance of the Gaussian mixtures included in the silence model so that the likelihood score would be smoothed for various inputs.

Figure 8 shows the results obtained by multiplying 1.5 to all diagonal elements of the covariance matrix of the Gaussian mixtures included in the silence model. After introducing such a simple modification, the drop of the recognition rate near \( \alpha = 0.3 \) in Fig. 6 disappeared, and higher recognition rates were obtained with \( \alpha \) larger than 0.2. The only drawback of this approach is that the baseline recognition rate is slightly degraded (see recognition rates at \( \alpha = 0 \) in Figs. 6 and 8). In Fig. 8, the shape of the curve is now similar to those in the previous cases, and the best recognition rate of 83.2% was obtained by “bidirectional look-ahead” with \( \alpha = 0.3 \). This indicates that the influence of musical noise in the speech period is not as fatal as in the nonspeech period.

5. CONCLUSIONS

In this paper, we proposed a new noise-suppression method based on the look-ahead mechanism and bidirectional estimation. The proposed method is an expansion of the well-known OM-LSA speech estimator, but the newly introduced functions contribute to avoiding incorrect estimation near sudden changes in the speech and error accumulation over time.

We evaluated the proposed method using public and proprietary databases. The experiments using the CENSREC-2 public database revealed that the proposed method outperforms one of the best feature normalization methods under various conditions, but is less effective for only audio noise.

The experiments using our proprietary database showed similar results, but a problem arose when we added a time-
varying noise to the test utterances. The problem was attributed to the vulnerability of the silence model caused by musical noise in the nonspeech period. The recognition rate could be recovered when we enhanced the robustness of the silence model, and we expect that the speech period will be successfully reconstructed by the proposed method even if it is contaminated by the time-varying noise.

The proposed algorithm has the advantage that it can be used as a preprocessor for any speech recognition system. It is particularly advantageous for large-scale commercial systems, in which it is costly or impossible to modify the speech recognizer itself. We found that the proposed method is less effective for highly nonstationary noise, and only the modification of the speech recognizer can solve this problem. However, the modification only relates to the nonspeech period, and the combination of the proposed method and a state-of-the-art VAD is expected to provide equivalent performance while keeping its function as a pure preprocessor. Since some of the authors reported that the OM-LSA-based VAD is very accurate under noisy conditions [24], the pursuit of such improvement would be promising future work.

Latency control for longer inputs is another topic to be investigated. However, even though there are such issues to be examined in the future, it should again be emphasized that the proposed method will improve the accuracy of existing speech recognition systems merely by placing it between the speech acquisition module and the speech recognizer.

REFERENCES

[11] 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).
Yasunari Obuchi received the B.S. degree and the M.S. degree in physics in 1988 and 1990 respectively, and the Ph.D. in information science and technology in 2006, all from the University of Tokyo. Since 1992, he has been working for Central Research Laboratory and Advanced Research Laboratory, Hitachi Ltd. He worked at Carnegie Mellon University as a Visiting Researcher from 2002 to 2003. He had also been a Visiting Researcher at Waseda University from 2005 to 2010. Currently he is a Senior Researcher of Central Research Laboratory, Hitachi Ltd. His research interests include robust speech recognition, spoken dialog systems, speech recognition in small devices, speech-to-speech translation, language identification, emotion recognition, and microphone arrays. Dr. Obuchi was a co-recipient of the Technology Development Award of the Acoustical Society of Japan in 2000. He is a member of IEEE, IPSJ, and IEICE.

Ryu Takeda received the B.E., M.Inf., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 2006, 2008, and 2011, respectively. He jointed the Hitachi Central Research Laboratory in 2011. Since then, he has worked on speech recognition. His research interests include speech enhancement and machine learning.

Masahito Togami received a B.E., M.E. degree, and Ph.D. in Aerospace Engineering from the University of Tokyo, Japan, in 2000, 2002, and 2011, respectively. He joined the Central Research Laboratory, Hitachi Ltd. in Japan in 2003. He is currently a senior researcher in the Central Research Laboratory, Hitachi Ltd. His research interests are array signal processing, acoustic echo canceller and automatic speech recognition of robots. He received the Annual Conference Award from JSAI in 2003. He received the Good Design Award in 2008. He received the Awaysa Award, the Itakura Award, and the Poster Award all from ASJ in 2010. He is a member of IEEE, IEICE, and JSAI.