FAST SEGMENT SEARCH FOR CORPUS-BASED SPEECH ENHANCEMENT BASED ON SPEECH RECOGNITION TECHNOLOGY

Atsunori Ogawa, Keisuke Kinoshita, Takaaki Hori, Tomohiro Nakatani and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation
{ogawa.atsunori, kinoshita.k, hori.t, nakatani.tomohiro, nakamura.atsushi}@lab.ntt.co.jp

ABSTRACT

Corpus-based speech enhancement has received increasing attention recently since it shows high enhancement performance in highly non-stationary noisy environments by precisely modeling the long-term temporal dynamics of speech. However, it has a disadvantage in that the cost is very high for searching the longest matching clean speech segments from a multi-condition parallel speech corpus. This paper proposes a *fast segment search method for corpus-based speech enhancement*. It is mainly based on two techniques derived from speech recognition technology. The first is an A* search like segment evaluation function for accurately finding the longest matching segments. The second is a tree and linear connected search space for efficiently sharing the segment likelihood calculations. In the experiments for non-stationary noisy observations using the 26 multi-condition TIMIT parallel speech corpus, the proposed search method found the segments almost in real-time without degrading the quality of the enhanced speech. Our method was about 7 to 13 times faster than the conventional segment search method.

Index Terms— Corpus-based speech enhancement, fast segment search, longest matching segments, speech recognition technology

1. INTRODUCTION

Speech enhancement is an essential technology for significantly improving the quality of speech-based applications, e.g. conversations over mobile phones and voice command inputs to car navigation systems, in adverse environments. And a lot of effort has been expended over many years on developing various types of effective speech enhancement approaches [1]. Among them, single-channel approaches have been the most actively studied, e.g. [1–14]. And they can be categorized primarily into two types with different advantages and disadvantages.

The first category consists of common and widely used filtering-based approaches that estimate noise statistics, typically, the noise power spectral density (PSD), e.g. [1–8]. And the estimated statistics are used to filter out the noise component from a noisy observation. If the noise is stationary, it is possible to estimate its statistics using the non-speech period of the observation (e.g. at the beginning of a recording session) [1, 2]. However, in many applications, the noise statistics can vary over time, and thus they must be tracked continuously. Many noise tracking approaches have been proposed, e.g. the minimum statistics approach [3, 4], the minima-controlled recursive averaging [5, 6], the minimum mean-square error based noise PSD estimator [7], and the recursive expectation-maximization algorithm [8]. The advantage of these approaches is their reasonable computational complexity. The accuracy of the noise tracking, and thus the quality of the enhanced speech, has been steadily improved with these approaches. However, the tracking of highly non-stationary noise remains as a very difficult task.

The second category consists of the recently proposed *corpus-based* (e.g. [9, 10]) and *inventory-style* (e.g. [11, 12]) approaches. In contrast to the approaches in the first category, these approaches focus strongly, not on the estimation of the noise statistics, but on the direct estimation of the underlying clean speech component. For example, the corpus-based approach proposed in [9,10] can be outlined as follows. First, a multi-condition parallel speech corpus and the corresponding *segment models* are prepared to capture the precise long-term temporal dynamics of speech. Then, given a noisy observation, using the segment models, the *longest matching* clean speech segments to the noisy input are found. Finally, by concatenating the found segments, clean speech is resynthesized. By compensating for the noise component in a noisy observation using the multi-condition parallel speech corpus, the approach can focus solely on the estimation of the underlying clean speech component. In addition, based on the longest matching property in the segment search, it can robustly find the correct clean speech segments that match the noisy input. In fact, this approach has exhibited high enhancement performance in highly non-stationary noisy environments [9, 10]. However, these types of approaches have an obvious disadvantage in that the cost of searching for the matching segments is very high [13, 14].

In this paper, we propose a *fast segment search method for corpus-based speech enhancement*. We assume the framework proposed in [9, 10] (described in Section 2.1). In these references, however, there is only a limited discussion of the acceleration of the segment search and their effect is not revealed concretely. In [9, 10], a segment evaluation function is defined that has the longest matching property (Section 2.2). In contrast, we propose another segment evaluation function, which is derived from the *A* search* technique in *speech recognition* technology and also has the longest matching property (Section 3.1). It is simpler and mathematically more rigorous than the conventional function and can find the segments accurately. In [9, 10], the segment search is conducted in an unstructured search space. In contrast, we introduce a *tree and linear connected search space* that is derived from the *tree lexicon* also employed in speech recognition (Section 3.2). Based on an analysis of the segment variations, the search space is designed to efficiently represent the speech corpus (i.e. the collection of segments). And by using this *structured* search space, the segment likelihood calculations can be efficiently shared. In the experiments for non-stationary noisy observations using the 26 multi-condition TIMIT parallel speech corpus, the proposed search method found the segments almost in real-time without degrading the quality of the enhanced speech (Section 4). Our method was about 7 to 13 times faster than the conventional segment search method.

2. CORPUS-BASED SPEECH ENHANCEMENT

This section briefly describes the basic framework of the corpus-based approach proposed in [9, 10] using Fig. 1 and details its conventional segment search method.

2.1. Basic Framework

In the training stage (top dotted box in Fig. 1), a clean speech corpus is first prepared. It is artificially contaminated with various types

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

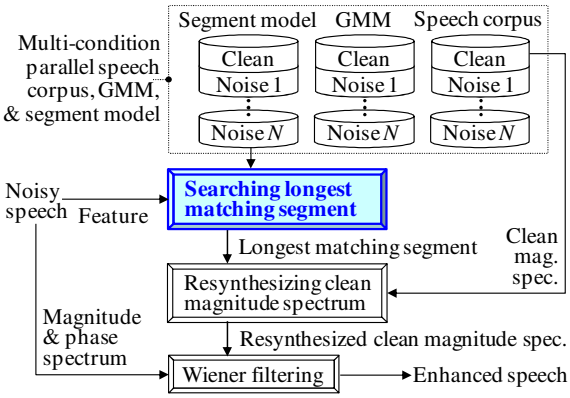

Fig. 1. Basic framework of corpus-based speech enhancement.

of noise to form a multi-condition parallel speech corpus. Feature (e.g. Mel-Frequency Cepstral Coefficient: MFCC) extraction is conducted for all of the speech corpora (as for the clean corpus, the magnitude spectra are also extracted). Using the extracted features, GMMs that represent each of the corpora are trained. To represent the precise spectral patterns of speech, the number of Gaussian components in each GMM is set large (e.g. 4096). And using these GMMs, *segment models* can be obtained (detailed in Section 2.2).

Then, given an input noisy speech, we first extract its feature, magnitude spectrum and phase spectrum sequences. Using the segment models, we can find the *longest matching segment* sequence to the input noisy speech with the segment posterior probabilities. The noise component in the noisy input can be compensated by using the multi-condition parallel corpus. However, the computational cost of this segment search is very high (detailed in Section 2.2).

Using the found matching segment sequence and the segment posterior probabilities, we resynthesize a clean magnitude spectrum sequence by concatenating the corresponding clean speech magnitude spectra. Finally, we perform Wiener filtering using the resynthesized clean magnitude spectrum sequence and the magnitude and phase spectrum sequences extracted from the input noisy speech to obtain the final enhanced speech.

2.2. Conventional Segment Search Method

Hereafter, for simplicity, we assume a single speech corpus (the experiments in Section 4 are conducted using a multi-condition parallel corpus) and that all the utterances in the corpus are concatenated into one long utterance. We employ the notations used in [9, 10].

We start with the training stage. Let $\mathbf{x} = \{x_i : i = 1, 2, \ldots, I\}$ be the whole feature sequence in a speech corpus, x_i be the feature at time frame i, and I be the total number of frames in the corpus. Using $\mathbf{x}$, a GMM G can be trained as

$$G = \{g(x|m), w(m) : m = 1, 2, \ldots, M\}, \tag{1}$$

where $g(x|m)$ is the m-th Gaussian component, $w(m)$ is the corresponding weight, and M is the total number of Gaussian components in G. Using G, we can obtain a model that represents the patterns of the temporal dynamics contained in the corpus $\mathbf{x}$. That is, for each time frame i, we find the Gaussian component m in G that maximizes the likelihood of the feature x_i. This results in a time sequence of maximum-likelihood Gaussian component indices as

$$\mathbf{m} = \{m_i : i = 1, 2, \ldots, I\}, \tag{2}$$

where m_i is an index addressing a Gaussian component $g(x|m_i)$ in G. $g(x|m_i)$ represents a class of short-time speech spectra, and thus, $\mathbf{m}$ can be used as a spectral-temporal model of the corpus $\mathbf{x}$. We refer to this model as the *segment model*.

Now, we describe the segment search. Let $\mathbf{y} = \{y_t : t = 1, 2, \ldots, T\}$ be a T frame feature sequence of a noisy speech input and y_t be the feature at time frame t. Hereafter, again for simplicity, we add a constraint that only allows the one-to-one frame basis linear matching (i.e. does not allow the elastic matching, e.g. dynamic programming (DP) or dynamic time warping) during the matching between $\mathbf{y}$ and $\mathbf{m}$ (see [9, 10] for details). With this constraint, let $\mathbf{y}_{t:t+\tau} = \{y_\epsilon : \epsilon = t, t+1, \ldots, t+\tau\}$ be an input segment taken from the time frames t to $t+\tau$ of $\mathbf{y}$ and $\mathbf{m}_{u:u+\tau} = \{m_i : i = u, u+1, \ldots, u+\tau\}$ be a corpus segment taken from time frames u to $u+\tau$ of $\mathbf{m}$. Then, at each time frame t in $\mathbf{y}$, we can find an input segment $\mathbf{y}_{t:t+\tau_{\max}}$ and the corresponding matching corpus segment $\mathbf{m}^t_{u:u+\tau_{\max}}$ by maximizing the posterior probability as

$$\mathbf{m}^t_{u:u+\tau_{\max}} = \arg\max_{\tau} \max_{\mathbf{m}_{u:u+\tau}} P(\mathbf{m}_{u:u+\tau}|\mathbf{y}_{t:t+\tau}), \tag{3}$$

$$P(\mathbf{m}_{u:u+\tau}|\mathbf{y}_{t:t+\tau}) = \frac{p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u:u+\tau})}{\sum_{u'} p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u':u'+\tau}) + p(\mathbf{y}_{t:t+\tau}|\phi_{t:t+\tau})}, \tag{4}$$

where $P(\mathbf{m}_{u:u+\tau}|\mathbf{y}_{t:t+\tau})$ is the posterior probability that has an important characteristic; It favors *longer matching*, i.e. a larger τ, between $\mathbf{y}_{t:t+\tau}$ and $\mathbf{m}_{u:u+\tau}$. So the longer the matched segment length is, the higher the posterior probability becomes (the proof is given in [10]). This property is important since longer speech segments can be identified more accurately in noisy environments than shorter segments because of their more distinct and richer spectral-temporal pattern information.

The numerator of Eq. (4) is the likelihood of the input segment $\mathbf{y}_{t:t+\tau}$ given the corpus segment $\mathbf{m}_{u:u+\tau}$ and is calculated as

$$p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u:u+\tau}) = \prod_{\epsilon=0}^{\tau} g(y_{t+\epsilon}|m_{u+\epsilon}), \tag{5}$$

where we assume the conditional independence of the adjacent frames. The denominator of Eq. (4) has two terms. The first term is a collection consisting of all the corpus segments, which are taken from all the possible segment locations (i.e. u'), and that are likely to match the given input segment $\mathbf{y}_{t:t+\tau}$. The second term is the likelihood of $\mathbf{y}_{t:t+\tau}$ given the corpus GMM G and is calculated as

$$p(\mathbf{y}_{t:t+\tau}|\phi_{t:t+\tau}) = \prod_{\epsilon=0}^{\tau} \left[\sum_{m=1}^{M} w(m) g(y_{t+\epsilon}|m) \right], \tag{6}$$

where the sum inside the brackets is the GMM-based likelihood for the feature $y_{t+\epsilon}$ (see [9, 10] for the meaning of this term).

In the implementation of the segment search at time frame t in $\mathbf{y}$ (i.e. the way to solve Eq. (3), top of Fig. 2), we first set $\tau = 0$ and find the most probable segment of length 1 in the corpus $\mathbf{x}$. We repeat this procedure while increasing τ by 1 until it reaches the previously determined maximum limit value $\tau_{\lim}$ (e.g. $\tau_{\lim} = 29$). After obtaining the most probable corpus segment for each τ, we find the maximum matching segment with $\tau_{\max}$ ($0 \le \tau_{\max} \le \tau_{\lim}$), that should result in the maximum posterior probability.

The conventional segment search method described above has two main problems. The first is that, when finding the maximum matching corpus segment, it compares the posterior probabilities between the candidate segments of *different lengths* (i.e. 1 to $\tau_{\lim}+1$). This comparison does not appear to be mathematically rigorous. In addition, the proof of the longest matching property given in [10] is rather complex. The second problem is purely concerned with the computational cost. The number of possible segment locations (i.e. u) in the corpus $\mathbf{x}$ is I, i.e. the total number of frames in the corpus. If I is large (and actually it is large), it is obvious that the cost of likelihood calculations for all the possible locations becomes very high (despite introducing the caching of once calculated Gaussian component likelihoods at each time frame t in $\mathbf{y}$). In [10], the pruning of unlikely segment hypotheses (locations) is introduced while increasing τ. Nevertheless, at least for the first frame (i.e. $\tau = 0$) of each search, all the I possible locations must be considered.

3. PROPOSED SEGMENT SEARCH METHOD

To address the two problems with the conventional segment search method, we propose the following two techniques that are derived from *speech recognition* technology.

3.1. A* Search Like Segment Evaluation Function

To address the first problem, we propose to evaluate the input segment $\mathbf{y}_{t:t+\tau}$ by dividing it into two sub-segments. The first sub-segment $\mathbf{y}_{t:t+\nu}$ ($0 \leq \nu \leq \tau$) is evaluated by the corpus segment $\mathbf{m}_{u:u+\nu}$ and the second sub-segment $\mathbf{y}_{t+\nu+1:t+\tau}$ is evaluated by the corpus GMM G (denoted as $\phi_{u+\nu+1:u+\tau}$). By assuming the conditional independence of the adjacent frames, this segment evaluation function (i.e. likelihood) can be written as

$$\begin{aligned} & p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u:u+\nu}, \phi_{u+\nu+1:u+\tau}) \\ &= p(\mathbf{y}_{t:t+\nu}|\mathbf{m}_{u:u+\nu}) p(\mathbf{y}_{t+\nu+1:t+\tau}|\phi_{u+\nu+1:u+\tau}) \qquad (7) \\ &= \prod_{\epsilon=0}^{\nu} g(y_{t+\epsilon}|m_{u+\epsilon}) \prod_{\epsilon=\nu+1}^{\tau} \left[\sum_{m=1}^{M} w(m) g(y_{t+\epsilon}|m) \right]. \qquad (8) \end{aligned}$$

The definition of this function is inspired by the A^* *search* technique in speech recognition, e.g. [15, 16], where a hypothesis is evaluated with the likelihood of an already searched segment calculated using a precise model *plus* the likelihood of an unsearched future segment estimated using a rough model. This is undertaken to obtain an *equal* comparison of hypotheses with different length searched segments on the basis of the common whole segment length, i.e. the input utterance length. Using this A^* *search like segment evaluation function*, at each time frame t in $\mathbf{y}$, we can find an input segment $\mathbf{y}_{t:t+\nu_{\max}}$ and the corresponding matching corpus segment $\mathbf{m}^t_{u:u+\nu_{\max}}$ by maximizing the posterior probability as

$$\mathbf{m}^t_{u:u+\nu_{\max}} = \arg\max_{\nu} \max_{\mathbf{m}_{u:u+\nu}} P(\mathbf{m}_{u:u+\nu}, \phi_{u+\nu+1:u+\tau}|\mathbf{y}_{t:t+\tau}), \qquad (9)$$

$$\begin{aligned} & P(\mathbf{m}_{u:u+\nu}, \phi_{u+\nu+1:u+\tau}|\mathbf{y}_{t:t+\tau}) \\ &= \frac{p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u:u+\nu}, \phi_{u+\nu+1:u+\tau})}{\sum_{u'} \sum_{\nu'} p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u':u'+\nu'}, \phi_{u'+\nu'+1:u'+\tau})}, \qquad (10) \end{aligned}$$

where the denominator of Eq. (10) is the sum of the segment likelihoods corresponding to all the possible corpus segment locations (i.e. u') and all the possible segment division boundaries (i.e. ν').

Note that, in the implementation of Eq. (9) (bottom of Fig. 2), by fixing the length of an input segment $\mathbf{y}_{t:t+\tau}$ at $\tau_{\lim}+1$ (i.e. the maximum limit length) the posterior probabilities of all the possible corpus segments $\mathbf{m}_{u:u+\nu}$ of different lengths (i.e. 1 to $\tau_{\lim}+1$) are compared *equally* on the basis of the common length, i.e. $\tau_{\lim}+1$. And this comparison seems to be mathematically rigorous.

The posterior probability of Eq. (10) has the longest matching property as with that of Eq. (4). The proof is very simple as follows: We compare the two posterior probabilities of the input segment $\mathbf{y}_{t:t+\tau}$; one is evaluated with the corpus segment $\mathbf{m}_{u:u+\nu}$ and the corpus GMM $\phi_{u+\nu+1:u+\tau}$ and the other is evaluated with $\mathbf{m}_{u:u+\nu-1}$ and $\phi_{u+\nu:u+\tau}$. The denominator is common to both probabilities and their ratio is equal to the likelihood ratio as

$$\frac{p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u:u+\nu}, \phi_{u+\nu+1:u+\tau})}{p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{u:u+\nu-1}, \phi_{u+\nu:u+\tau})} = \frac{g(y_{t+\nu}|m_{u+\nu})}{\sum_{m=1}^{M} w(m) g(y_{t+\nu}|m)}. \qquad (11)$$

Here, we assume that the whole acoustic space is equally covered by each of the Gaussian components in G and the feature $y_{t+\nu}$ is well-matched to the Gaussian component $m_{u+\nu}$. With these two assumptions, the denominator on the right-hand side of Eq. (11) can be approximated as $w(m_{u+\nu})g(y_{t+\nu}|m_{u+\nu})$, and thus, Eq. (11) becomes equal to $1/w(m_{u+\nu}) \geq 1$. This means that, as long as there is a Gaussian component $m_{u+\nu}$ that matches the feature $y_{t+\nu}$, the matching corpus segment $\mathbf{m}_{u:u+\nu}$ becomes long.

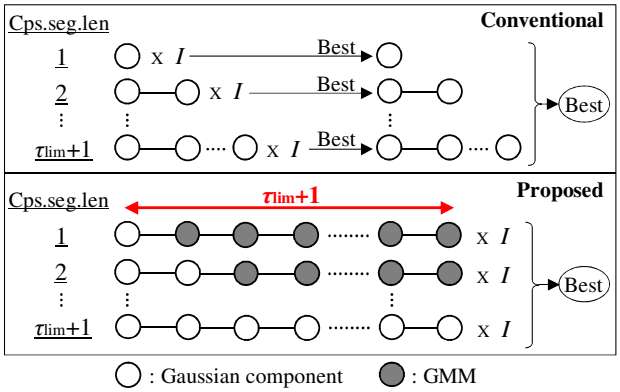

Fig. 2. (Top) Conventional and (bottom) proposed corpus segment search procedures for an input segment.

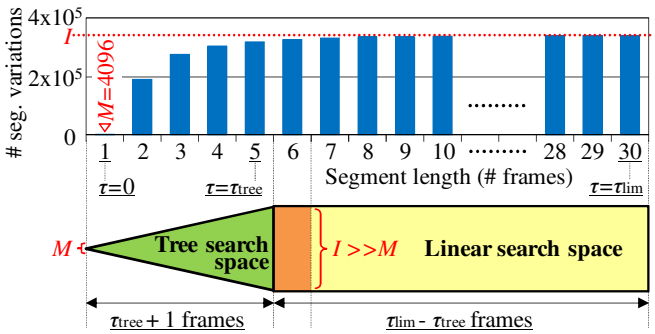

Fig. 3. (Top) Number of corpus segment variations as a function of segment length in a speech corpus and (bottom) proposed tree and linear connected search space.

3.2. Tree and Linear Connected Search Space

To address the second problem, we first counted the variations of corpus segment $\mathbf{m}_{t:t+\tau}$ as a function of segment length (i.e. $0 \leq \tau \leq \tau_{\lim}$) using a speech corpus. The result is shown in the top of Fig. 3. It is obvious that, when the segment length is 1 (i.e. $\tau = 0$), the number of segment variations is equal to M ($= 4096$), i.e. the total number of Gaussian components in the corpus GMM G. Then, it rapidly increases and, within a few frames, it converges to its upper bound I, i.e. the total number of frames in the corpus.

This result indicates that a *tree structure* can be introduced for the first few frames of the corpus segments (e.g. $0 \leq \tau \leq 4 = \tau_{\text{tree}}$) as shown in the bottom of Fig. 3. This is inspired by the *tree lexicon* in speech recognition, e.g. [17, 18]. A corpus segment and its Gaussian component sequence correspond to a word and its phoneme sequence (i.e. pronunciation). Using a *tree search space*, the likelihood calculations can be efficiently *shared* between the possible segments. In particular, the number of likelihood calculations for the first frame (i.e. $\tau = 0$) can be greatly reduced, i.e. from I with the conventional method (Section 2.2) to M ($M \ll I$).

For the remaining frames (i.e. $\tau_{\text{tree}} < \tau \leq \tau_{\lim}$), we introduce a *linear search space* as shown in Fig. 3. And by connecting it to the tree search space, we can obtain the proposed *tree and linear connected search space*. This connection is important in terms of memory usage, because the tree search space requires a memory size of $\tau_{\text{tree}}+1$ frames (i.e. the green area in Fig. 3), in contrast, the linear search space can be represented with only a one frame memory size (i.e. the orange area). By introducing the linear structure, the total memory requirement can be reduced to less than 10% compared with that when the whole search space is represented with only the tree structure. Of course, as with [10], the pruning of unlikely segment hypotheses can be introduced to our search space.

4. EXPERIMENTS

We compared the proposed segment search method with the conventional one experimentally in non-stationary noisy environments.

4.1. Experimental Setup

The training condition was basically the same as in [9,10]. The clean training data was the TIMIT speech corpus, which consists of 1088 utterances from 136 female speakers. These utterances were sampled at 16 kHz. The corpus size was about 56 minutes, i.e. the total number of frames was about 3.4×10^5 ($= I$). This clean corpus was artificially contaminated with 25 different conditions of white noise (see [9, 10] for details) to form a 26 multi-condition parallel speech corpus. The frame length and shift were 20 and 10 ms, respectively, and the feature was a 39-dimensional MFCC with a log energy term. Using the parallel corpus, the corresponding 26 GMMs were trained. Each GMM had 4096 ($= M$) Gaussian components.

For testing, we used the TIMIT *core* test set, which consists of 64 utterances from 8 female speakers. These clean utterances were contaminated with four different non-stationary noisy conditions: airport and factory noises with 0 and 5 dB signal-to-noise ratio (SNR) levels. In the segment search, the maximum limit length of a segment was set at 30 (i.e. $\tau_{\text{lim}} = 29$). The number of layers (frames) of the tree search space was set at 5 (i.e. $\tau_{\text{tree}} = 4$, only for the proposed method). The elastic segment matching was not allowed. And the pruning of unlikely hypotheses was introduced in the log-likelihood domain. The pruning beam width was set at 1, 2, 5, 10, 20 and 50. When a wider (narrower) beam width was set, the lengths of the matching segments became longer (shorter). The quality of the enhanced speech was measured by segmental SNR. And to confirm the quality of the corpus-based approach, as with [9, 10], Wiener filtering with *a priori* SNR [19], i.e. an approach that estimates the noise statistics (Section 1), was also evaluated. Both the segment search methods were implemented in C and run on a Linux system with Intel Xeon CPU X5690 3.47GHz. And the search speed was measured with a real time factor (RTF).

4.2. Experimental Results

Figure 4 shows the experimental results. From Fig. 4 (a) and (b), we can first confirm that the corpus-based approach significantly improves the quality of the enhanced speech compared with Wiener filtering as reported in [9,10]. We can also confirm that the best quality for both noise types is provided by the proposed segment search method. These results are thanks to the mathematical rigor of the proposed A* search like segment evaluation function. However, the difference is not very large between the quality of the conventional and proposed methods. Moreover, there is not a large difference between the quality of any of the pruning beam widths. However, the airport noise tends to favor wider beam widths (i.e. longer segments), in contrast, the factory noise tends to favor narrower beam widths (i.e. shorter segments). With the airport noise, the average segment length providing the best quality is about 8 frames. This is slightly shorter than that reported in [10], i.e. 11 frames. We think this is because we did not introduce the elastic segment matching that can provide longer segments. With the factory noise, the best average segment length is about 4 frames. We think this shorter length is caused by the characteristics of the factory noise, i.e. a quickly repeating impulsive noise. And to compensate for these characteristics, shorter segments are preferred.

From Fig. 4 (c), we can confirm that, for both the conventional and proposed methods, hypothesis pruning is very effective in accelerating the segment search without degrading the quality of the enhanced speech. We can also confirm that the proposed method can find the segments almost in real-time. It is about 7 to 13 times faster than the conventional method. This acceleration is obtained mainly by introducing the tree and linear connected search space. It is obvious that the proposed method gains a further advantage over the conventional method in that the corpus size (i.e. I) becomes larger since the sharing rate of the likelihood calculations (especially, for the first frame) becomes higher.

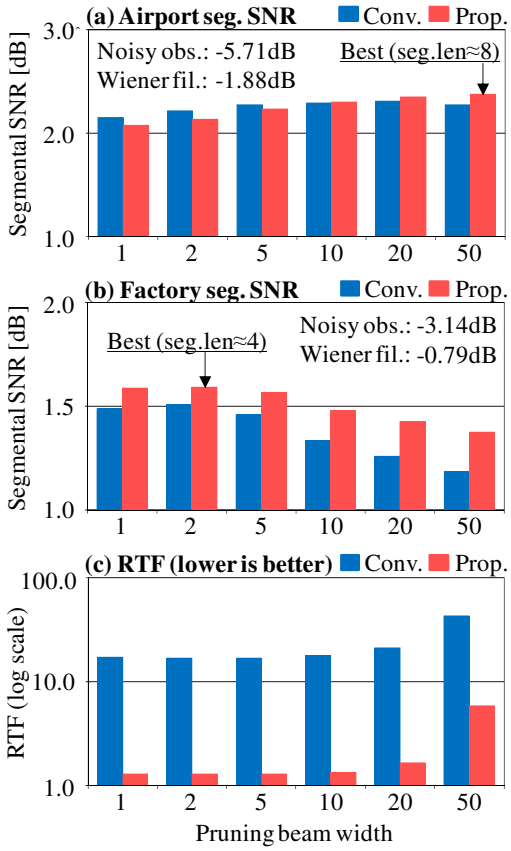

Fig. 4. (a) Segmental SNRs averaged over two SNR levels for airport noise, (b) those for factory noise, (c) RTFs averaged over all four noise conditions, as a function of pruning beam width.

5. RELATION TO PRIOR WORK

In [13], the optimal interconnection of a filtering-based approach and the corpus-based approach is proposed. In this connected method, the filtering-based approach acts as a preprocessor that reduces the noise component in noisy observations. As a result, the number of noise conditions in the parallel corpus can be reduced, thereby reducing the cost of the segment search.

In [14], an effective method is proposed for reducing the memory requirement and the computational complexity of the inventory-style approach. The memory reduction is accomplished with a singular value decomposition of the inventory matrices. And, using the decomposed matrices, a fast hierarchical sub-search (not an exhaustive full search) of the matching inventories can be performed.

Our fast segment search method is essentially different from both these methods. And so, we can expect to combine our method with these methods for further acceleration of the segment search.

6. CONCLUSION AND FUTURE WORK

We have proposed a fast segment search method for corpus-based speech enhancement based on speech recognition technology. In experiments, the proposed method found the segments almost in real-time without degrading the quality of the enhanced speech. Our method was about 7 to 13 times faster than the conventional method. Future work will include the incorporation of the elastic (e.g. DP) segment matching [9, 10] for quality improvement and combination with the methods proposed in [13, 14] for further acceleration.

7. REFERENCES

[1] P.C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*, CRC Press, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[3] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. Eurospeech*. ISCA, 1993, pp. 1093–1096.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[5] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[7] R.C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. ICASSP*. IEEE, 2010, pp. 4266–4269.

[8] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani, "Noise power spectral density tracking: A maximum likelihood perspective," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 495–498, August 2012.

[9] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," in *Proc. Interspeech*. ISCA, 2010, pp. 1097–1100.

[10] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822–836, May 2011.

[11] X. Xiao and R.M. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1243–1257, August 2010.

[12] R.M. Nickel, R.F. Astudillo, D. Kolossa, and R. Martin, "Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 983–997, May 2013.

[13] K. Kinoshita, M. Souden, M. Delcroix, and T. Nakatani, "Single channel dereverberation using example-based speech enhencement with uncertainty decoding technique," in *Proc. Interspeech*. ISCA, 2011, pp. 197–200.

[14] R.M. Nickel and R. Martin, "Memory and complexity reduction for inventory-style speech enhancement systems," in *Proc. EUSIPCO*. EURASIP, 2011, pp. 196–200.

[15] D.B. Paul, "New developments in the Lincoln stack-decoder based large-vocabulary CSR system," in *Proc. ICASSP*. IEEE, 1995, pp. 45–48.

[16] P.S. Gopalakrishnan, L.R. Bahl, and R.L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," in *Proc. ICASSP*. IEEE, 1995, pp. 572–575.

[17] R. Heab-Umbach and H. Ney, "Improvements in beam search for 10000-word continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 353–356, April 1994.

[18] S. Ortmanns, A. Eiden, H. Ney, and N. Coenen, "Look-ahead techniques for fast beam search," in *Proc. ICASSP*. IEEE, 1997, pp. 1783–1786.

[19] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*. IEEE, 1996, pp. 629–632.