SIP (2015), vol. 4, e3, page 1 of 12 © The Authors, 2015.

doi:10.1017/ATSIP.2015.5

ORIGINAL PAPER

Optimized wavelet-domain filtering under noisy and reverberant conditions

RANDY GOMEZ[1], TATSUYA KAWAHARA[2], AND KAZUHRIO NAKADAI[1]

The paper addresses a robust wavelet-based speech enhancement for automatic speech recognition in reverberant and noisy conditions. We propose a novel scheme in improving the speech, late reflection, and noise power estimates from the observed contaminated signal. The improved estimates are used to calculate the Wiener gain in filtering the late reflections and additive noise. In the proposed scheme, optimization of the wavelet family and its parameters is conducted using an acoustic model (AM). In the offline mode, the optimal wavelet family is selected separately for the speech, late reflections, and background noise based on the AM likelihood. Then, the parameters of the selected wavelet family are optimized specifically for each signal subspace. As a result we can use a wavelet sensitive to the speech, late reflection, and the additive noise, which can independently and accurately estimate these signals directly from an observed contaminated signal. For speech recognition, the most suitable wavelet is identified from the pre-stored wavelets, and wavelet-domain filtering is conducted to the noisy and reverberant speech signal. Experimental evaluations using real reverberant data demonstrate the effectiveness and robustness of the proposed method.

Keywords: Automatic speech recognition, Dereverberation, Robustness

Received 2 December 2014; revised 7 June 2015

I. INTRODUCTION

In a real-world enclosed environment, the speech signal is reflected and arrives at different time delays when observed by the microphone. This effect is considered as a form of a contamination due to channel distortion, and commonly known as reverberation. The degree of reverberation depends on the reverberation time T_{60}, which dictates the severity of distortion. Speech contamination is one of the most common problems in automatic speech recognition (ASR) applications. In the perspective of ASR, any form of contamination of the speech signal at runtime (test condition) is a mismatch to the acoustic model (AM) (training condition). The mismatch may result in the degradation of the ASR performance. Thus, speech enhancement is one of the most important topics in robust ASR. In this paper, we focus primarily on the topic of dereverberation for ASR; since background noise is always present in a real environment, we address enhancement in reverberant and noisy condition, and extend our dereverberation framework to include denoising effect.

The scheme of decomposition of the reverberant signal into early and late reflections [1] simplifies the treatment of reverberation. In this scheme, the late reflection is treated as additive noise, and the seminal works [2–5] in denoising has been adopted. We expanded multi-band spectral subtraction (SS) so that the multi-band weighting parameters are optimized based on the criterion of the ASR [6]. Similarly, the Wiener filtering (WF) approach can be employed to the same dereverberation scheme. Originally adopted from athe denoising work in [7], it can be extended to suppress the late reflection by filtering the reverberant signal with the Wiener gain. Although the filtering-based methods (e.g. SS and Wiener) work well, they share a common problem: power estimation of the contaminant (i.e. late reflection and background noise) and the desired signal (i.e. speech). This problem is inherent to the filtering-based methods [2–5]. In real scenario, both the contaminant and the speech signals are not available separately, instead, we need to deal with a composite signal, and extracting independent power estimates for each of these is not simple. Since the filtering-based methods depend primarily on power estimation, inaccurate estimates result in artifacts in the recovered signal. This impacts the ASR performance in general, as it manifests as another form of mismatch to the recognizer. Power estimation is improved through popular methods in the seminal works [8–11] coupled with the deployment of voice activity detector (VAD).

In this paper, we address the problem through optimal wavelet-domain filtering. WF is adopted as the enhancement platform, but instead of operating in the frequency domain, we perform filtering in the wavelet domain. Wavelets offer more flexibility in signal representation. A

[1]Honda Research Institute Co., Ltd., Wako-shi, Saitama 351-0188, Japan
[2]Kyoto University, ACCMS, Sakyo-ku, Kyoto 606-8501, Japan

Corresponding author:
R. Gomez
Email: r.gomez@jp.honda-ri.com

proper choice of wavelet allows us to track the power of the signal of interest directly from the observed contaminated signal. This mechanism results to a more accurate instantaneous (frame-wise) power estimates. This is not possible using traditional VAD relying on a priori information regarding speech/non-speech frames. Specifically, in this paper, we present a method in optimizing the wavelets based on the ASR criterion. We note that the ASR is a complex system and operates independently from speech enhancement (i.e. dereverberation) module. By setting the optimization criterion used in the dereverberation as a function of the AM used by the ASR, we can expect that the dereverberation method is optimized to improving ASR performance. Our previous work in [12, 13] addressed a very limited optimization which only covers wavelet parameter tuning. In the proposed method, optimization is more comprehensive via wavelet family selection and parameter optimization not covered in [12, 13]. Thus, optimization of the wavelet family and parameters is conducted using AM likelihood (Section III) for each signal of interest (i.e. speech, late reflection, and additive background noise). Since characteristics of these signals are different, optimizing the wavelet for each corresponding signal improves signal representation and power estimation for effective speech enhancement in ASR.

The paper is organized as follows. In Section II we present the enhancement concept of our proposed method which includes the formulation of the reverberant model and its expansion to include background noise, theory of WF in the wavelet domain, and the synergy between enhancement and the ASR system. Section III describes the optimization via wavelet family selection and wavelet parameter tuning. Then, the method of estimating the reverberation time T_{60} and the identification of the noise profiles using Gaussian mixture model (GMM) is discussed in Section IV. The experimental setup and ASR performance are presented in Section V. Finally, we conclude the paper in Section VI.

II. ENHANCEMENT CONCEPT

In this section, we present the concept of our enhancement approach by introducing the reverberant model we adopted from [1]. The formulation of the Wiener filter in the frequency domain together with its expansion to the wavelet domain is presented. Lastly, the wavelet optimization based on acoustic likelihood criterion is discussed.

A) Model for dereverberation

The reverberant spectra $R(f,w)$ (short-term spectrum, w: window frame, f: frequency) is given as

$$R(f,w) \approx S(f,w)H(f,w), \tag{1}$$

where $S(f,w)$ and $H(f,w)$ are the clean speech signal and the room impulse response (RIR), respectively. The RIR h can be expressed with early h_E and late h_L components of the RIR as follows:

$$h_E(t) = \begin{cases} h(t) & t < \Gamma, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

$$h_L(t) = \begin{cases} h(t+\Gamma) & t \geq \Gamma, \\ 0 & \text{otherwise}. \end{cases} \tag{3}$$

Equations (2) and (3) characterize both the short and long-period effects of the reverberant signal. From equation (1), the reverberant speech model $R(f,w)$ is expressed as

$$\begin{aligned} R(f,w) &\approx E(f,w) + L(f,w) \\ &\approx S(f,w)H(f,0) + \sum_{d=1}^{D} S(f,w-d)H(f,d). \end{aligned} \tag{4}$$

The first term is referred to as the early reflection denoted as $E(f,w)$, where $H(f,0)$ is the RIR effect to the speech signal $S(f,w)$. It is due to the direct-path signal contaminated with some reflections that occur at earlier time (short period). The second term $L(f,w)$, is attributed by late reflection, which can be viewed as smearing of the clean speech by $H(f,d)$ which corresponds to the d frame-shift effect of the RIR. D is the number of frames over which the reverberation (smearing) has an effect. Since the late reflection spans over frames, it can be treated as long-period noise [14, 15]. In real environments, it is safe to assume that some additive noise may be present. Although our main focus in this paper is about dereverberation, we include a simple additive noise mitigation scheme, since we experiment using real data and the presence of noise impacts the dereverberation mechanism in [6, 14].

In general, removing contaminants especially the effects of late reflection and background noise is a difficult task. Since the dereverberation concept in this paper was originally inspired from denoising [2–5], the treatment of noise together with the effects of reverberation is possible as long as the following assumptions are adopted:

- Decomposition of reverberation into early and late reflection.
- Additive property of late reflection and noise.
- Statistical independence and uncorrelatedness of the signals (i.e. speech, late reflection, and additive noise).

Following equation (4), we can include the effects of the additive background noise $B(f,w)$ and the observed contaminated signal $O(f,w)$ becomes

$$\begin{aligned} O(f,w) &\approx R(f,w) + B(f,w) \\ &\approx E(f,w) + L(f,w) + B(f,w). \end{aligned} \tag{5}$$

In equation (5), we assume that the early reflection, late reflection and background noise are uncorrelated and statistically independent. However, this assumption may not hold true; thus, we show later an optimization process aimed to further strengthen the assumption in the wavelet

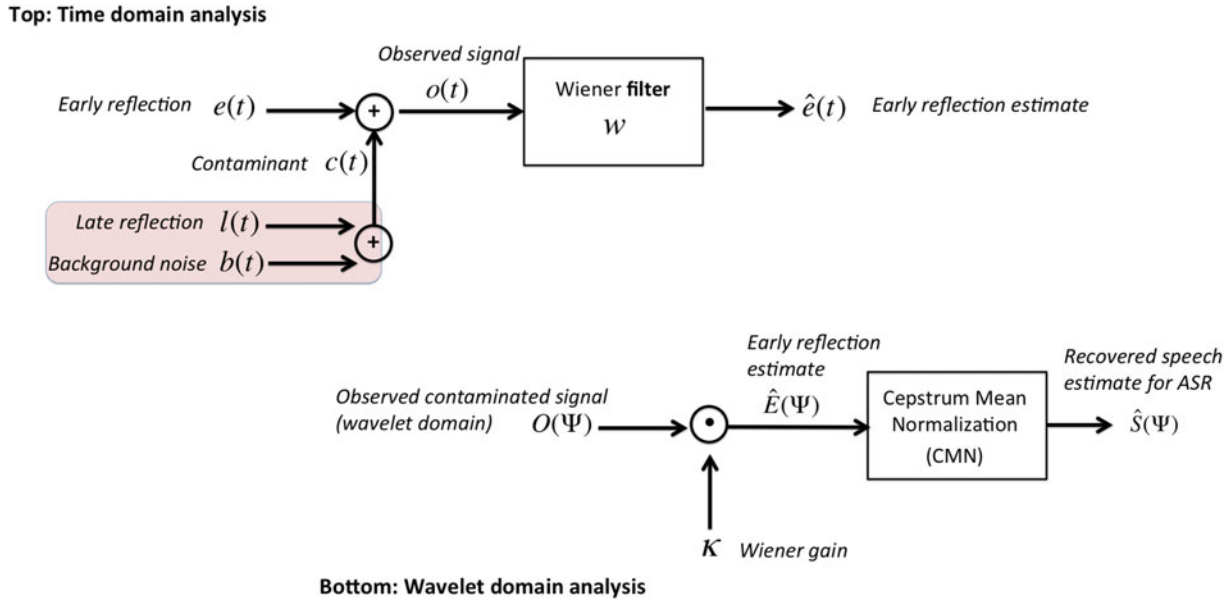

Fig. 1. Overview of the enhancement model.

domain. From here, we refer to the combined effects of the late reflection and the background noise as contaminant. Speech is enhanced by suppressing $L(f,w)$ and $B(f,w)$. Consequently, the recovered early reflection is processed via Cepstral Mean Normalization (CMN) [6] prior to the ASR. From this point onward, we assume that processing is conducted in framewise manner, dropping the index w.

B) WF in the wavelet domain

WF is an enhancement method based on the stochastic filter theory. Enhancement of the contaminated signal is based on the choice of the coefficients of the Wiener filter [16], and by imposing a criterion that minimizes the minimum mean square error (MMSE) between the desired and observed signals, the enhanced signal resembles that of the desired signal in the MMSE sense. Consider the conventional Wiener filter in Fig. 1 (top), the recovered early reflection $\hat{e}$ can be expressed as

$$\hat{e}(t) = o(t) * w(t), \tag{6}$$

where $w(t)$ is the Wiener filter impulse response. By applying the Wiener–Khinchine relation

$$E^2(f) = \mathcal{FT}\{R_{ee}(\nu)\}, \tag{7}$$

the autocorrelation function R_{ee} is replaced in terms of power spectra [7]. Thus, we can formulate the Wiener filter gain in terms of the frequency domain as

$$\begin{aligned} W(f) &= \frac{E^2(f)}{E^2(f)+C^2(f)} \\ &= \frac{E^2(f)}{E^2(f)+[L^2(f)+B^2(f)]}, \end{aligned} \tag{8}$$

where $E^2(f)$ and $C^2(f)$ are the power of early reflection and the contaminant, respectively. The contaminant signal is composed of the late reflection and background noise $(L^2(f)+B^2(f))$. By maximizing the discriminative property between the subspaces E and C, and by a proper wavelet choice Ψ (Section III), equation (8) can be expanded in the wavelet domain as

$$W(\Psi) = \frac{E^2(\Psi_E)}{E^2(\Psi_E)+[L^2(\Psi_L)+B^2(\Psi_B)]}, \tag{9}$$

where $E^2(\Psi_E)$, $L^2(\Psi_L)$, and $B^2(\Psi_B)$ are the early reflection, late reflection and background noise power, respectively. The wavelet domain Ψ_E, Ψ_L, and Ψ_B are the corresponding wavelets that enhances the discriminative property among subspaces E, L, and B (Section III). In the context of fast wavelet transform, we define κ from equation (9) at band m as

$$\kappa_m = \frac{E_m^2}{E_m^2+L_m^2+B_m^2}, \tag{10}$$

where band m denotes the level of the wavelet decomposition. In reality, we are interested in recovering the clean speech and not the early reflection. Although these two are not strictly the same (i.e. waveform-wise), they are almost the "same" as far as the ASR is concerned. We have concurred through experiments in [6] that the ASR is robust to the early reflection when processed with CMN; interestingly, recognition performance for CMN-processed signal using clean AM is comparable with that of clean speech signal. This means that the ASR makes no distinction between clean speech and CMN-processed early reflection. The CMN is able to compensate the mismatch caused by short-term smearing of the speech signal. By exploiting this behavior of the ASR, we can replace $E^2(\Psi_E)$ with $S^2(\Psi_S)$ in

equation (9). Thus, we can rewrite equations (9) and (10) as

$$W(\Psi) = \frac{S^2(\Psi_S)}{S^2(\Psi_S) + L^2(\Psi_L) + B^2(\Psi_B)} \tag{11}$$

and

$$\kappa_m = \frac{S_m^2}{S_m^2 + L_m^2 + B_m^2}. \tag{12}$$

This assumption was also confirmed in [6, 14, 15]. The ASR-inspired WF is implemented in the wavelet domain after wavelet decomposition as shown in Fig. 1 (bottom). First, the early reflection is recovered by weighting the observed contaminated signal with Weiner gain as

$$\hat{E}_m = \kappa_m \cdot O_m, \tag{13}$$

then, the enhanced speech is recovered after CMN processing expressed as,

$$\hat{S}_m = \text{CMN}(\hat{E}_m). \tag{14}$$

The time-domain speech can be recovered through inverse wavelet transform (IWT). It is obvious that the enhancement ability of the system is dependent on the Wiener gain, which is a function of the power estimates of the signals of interest. Unfortunately, there is no straightforward solution to power estimation. The observed signal at the microphone is a composite signal, which makes it difficult to estimate the speech power and the contaminant power (late reflection and background noise) independently and accurately. Conventionally, a VAD is used to improve power estimation as presented in [7, 17]. The contaminant power is estimated from the non-speech parts of the utterance. With the presumption that contaminant frames are of low-power as opposed to the composite signal which includes the speech. The speech power can be estimated by subtracting the observed signal with the contaminant estimate. Although this method works, estimation tends to be inaccurate especially with shorter utterances that do not have sufficient non-speech frames. Moreover, since the VAD method needs several frames for improved power estimation performance, it is difficult to calculate instantaneous power at a particular frame, resulting to poor performance in tracking the contaminant signal.

Speech enhancement performance relies primarily on the effectiveness of the contaminant power estimate, specifically for filtering-based methods. It is imperative to address the power estimation problem. In this paper, the expansion of WF to the wavelet domain presents a viable alternative in improving the power estimation (Section III).

C) Optimization via AM criterion

Speech enhancement and ASR are independent and complex processes. Originally, speech enhancement was developed to suppress noise and improve speech intelligibility for human listening, later it has been adopted for robust ASR application. However, human and machine perceive speech differently, and by simply cascading these two processes may not be effective [18]. Improvement in perceptual objective or subjective measures does not necessarily translate to improving ASR performance. As mentioned earlier, enhancement also introduces artifacts which may be detrimental to model-based ASR systems. Since there are many factors affecting ASR performance, it is appropriate to design an ASR-inspired speech enhancement method, and adopt the ASR criterion in the enhancement process. One of the most important features of the proposed method is the intricate link between the enhancement process and the ASR system.

The basic hidden Markov model (HMM)-based ASR system employs AM λ and language model in decoding speech to a word sequence. λ is obtained usually by maximum likelihood estimation

$$\max \prod_{r=1}^{R} P(\mathbf{x}_r|\mathbf{w};\lambda), \tag{15}$$

where $\mathbf{w}$ is the word sequence. $\mathbf{x}_r$ is the rth training utterance. Since equation (15) is an integral part of an ASR system, it is desirable to adopt this in the speech enhancement process. In conjunction with equation (15), we define the likelihood criterion score

$$L(\mathbf{x},\lambda) = p(\mathbf{x}|\lambda), \tag{16}$$

to measure the similarity between the signal $\mathbf{x}$ and the AM λ. For speech enhancement, $\mathbf{x}$ becomes the enhanced speech while λ is the AM used by the ASR. For computational efficiency, we adopt a GMM instead of a HMM to compute the AM likelihood. The likelihood score L increases when there is a good match between $\mathbf{x}$ and λ. This is a potent measure that relates the enhanced speech with the AM used by the ASR system. We note that $\mathbf{w}$ is purposely removed in equation (16) since we are only interested in the acoustic part of speech enhancement. Equation (16) will be extensively used throughout this paper for the wavelet optimization process (i.e. in Section III) used in our proposed dereverberation scheme.

III. WAVELET OPTIMIZATION FOR ENHANCEMENT

In this section, we discuss the optimization of the wavelet family and parameters using AM, which is the major contribution of this paper. A wavelet is generally expressed as

$$\Psi(\upsilon,\tau,t) = \frac{1}{\sqrt{\upsilon}}\Psi\left(\frac{t-\tau}{\upsilon}\right), \tag{17}$$

where t denotes time, υ and τ are the scaling and shifting parameters, respectively. $\Psi([t-\tau]/\upsilon)$ is often referred to as a mother wavelet. Assuming that we deal with real-valued signal, the wavelet transform (WT) is defined as

$$F(\upsilon,\tau) = \int z(t)\Psi(\upsilon,\tau,t)dt, \tag{18}$$

where $F(\upsilon,\tau)$ is a wavelet coefficient and $z(t)$ is a time-domain function. With a proper choice of wavelet family

coupled with a training algorithm, τ and υ are optimized, so that the wavelet captures the characteristics of the signal of interest. The resulting wavelet is sensitive to detecting the presence of the said signal. Specifically, in the wavelet filtering method, we are interested in detecting the power of speech, noise, and late reflection, given the observed signal at the microphone. Thus, we optimize the wavelet to detect these signals separately.

A) Wavelet family selection

LIKELIHOOD CRITERION

There exist different types of wavelets, bearing different waveforms (e.g. Daubechies and Morlet) referred to as wavelet family (f), and it is desirable to find an optimal match between the signal of interest and the corresponding wavelet family. We note that a particular wavelet family has a unique characteristic (i.e. waveform and frequency response) and may not be appropriate to represent a particular signal of interest. For example, Daubechies wavelet has the property which is more suitable to represent a speech signal than representing a background noise. We develop a process to quantify the distinction of which wavelet family is best suited for the signal of interest. The process is achieved by selecting a wavelet family Ψ among ($f = 1 : F$) denoted as $\Psi^{(f=1:F)}$ that best represents the signal of interest for speech S, late reflection L, and background noise B using the likelihood criterion given as

$$\psi_S = \arg\max_f p(\mathbf{s}(\Psi^{(f)})|\lambda_S), \tag{19}$$

$$\psi_L = \arg\max_f p(\mathbf{l}(\Psi^{(f)})|\lambda_L), \tag{20}$$

and

$$\psi_B = \arg\max_f p(\mathbf{b}(\Psi^{(f)})|\lambda_B), \tag{21}$$

where $\mathbf{s}(\Psi^{(f)})$, $\mathbf{l}(\Psi^{(f)})$, and $\mathbf{b}(\Psi^{(f)})$ are the speech, late reflection, and background noise processed with the wavelet family $\Psi^{(f)}$. λ_S, λ_L, and λ_B are AMs trained from clean speech, late reflection, and background noise database, respectively, using Mel-frequency cepstrum coefficients (MFCC) features. Equations (19)–(21) calculate the likelihood scores for the clean speech, late reflection, and background noise when decomposed using different wavelet family ($f = 1 : F$) against the corresponding AM. Thus, the corresponding wavelet family that results to the best decomposition of the signal of interest is selected based on the likelihood criterion.

LIKELIHOOD RATIO CRITERION

The likelihood criterion in Section III-A searches for the correspondence between the signal of interest and the collection of wavelet families. In this subsection, we introduce another optimization criteria focusing on the late reflection and background noise, so they are better separated from the speech signal.

In particular we search for the corresponding wavelet that maximizes the likelihood ratio between the speech model λ_S and the corresponding signal, given as

$$\psi_L = \arg\max \frac{p(\mathbf{l}(\psi_L)|\lambda_L)}{p(\mathbf{l}(\psi_L)|\lambda_S)} \tag{22}$$

and

$$\psi_B = \arg\max \frac{p(\mathbf{b}(\psi_B)|\lambda_B)}{p(\mathbf{b}(\psi_B)|\lambda_S)}. \tag{23}$$

Then, equation (11) becomes

$$W(\psi) = \frac{S^2(\psi_S)}{S^2(\psi_S) + L^2(\psi_L) + B^2(\psi_B)}. \tag{24}$$

B) Wavelet parameter optimization

A wavelet family is characterized by its parameters (i.e. scaling υ and shifting τ) as described in equation (18) which can significantly impact its response. Thus, we perform wavelet parameter optimization after the optimized wavelet family selection discussed in Section III-A. Optimizing both υ and τ will further refine the optimization process. The process of optimizing the wavelet parameters for each corresponding signal in the form of training as shown in Fig. 2 is discussed as follows:

SPEECH

Figure 2 (Top) shows the process of tuning the wavelet parameters for the speech signal. To conform with the model in Fig. 1 (bottom), the clean speech estimate $\hat{s}$ is synthesized by generating the early reflection using the the early components h_E of the RIR. The RIR can be synthetically generated as described in [6]. The RIR can be set as a function of reverberation time $T_{60}^{(j)}$ based on room acoustics model in [19]. Thus, physical measurement inside the room is not required. The theory behind the use of synthetic RIR stems from the fact that the HMM's description of speech is of low resolution compared with the RIR, with respect to time and frequency. Thus, for ASR applications, it may be sufficient to use an approximation of it [20]. And, its effectiveness in ASR is verified in [6]. Moreover, we have devised a method to effectively identify the early h_E and late h_L reflections boundary of the RIR in [6]. Thus, for a particular reverberation time j (i.e. $T_{60}^{(j)}$); $h_E^{(j)}$ and $h_L^{(j)}$ are readily available.

The early reflection estimate $\hat{e}$ is generated using the speech database and the corresponding early reflection coefficient h_E of the RIR. Then, CMN is applied resulting to $\hat{s}$. Wavelet coefficients $\hat{S}(\upsilon, \tau)$ is extracted through equation (18) using the optimized wavelet family in Section III-A. Likelihood scores are computed using the clean speech AM λ_S, a GMM of 256 components. λ_S is a text-independent model which captures the statistical information of the speech subspace. A greedy search process is iterated by adjusting υ and τ. The corresponding $\upsilon = a$ and $\tau = \alpha$ that result in the highest score are selected. Since we are interested in the speech subspace in general, optimizing a single wavelet to capture the general speech characteristics is sufficient.

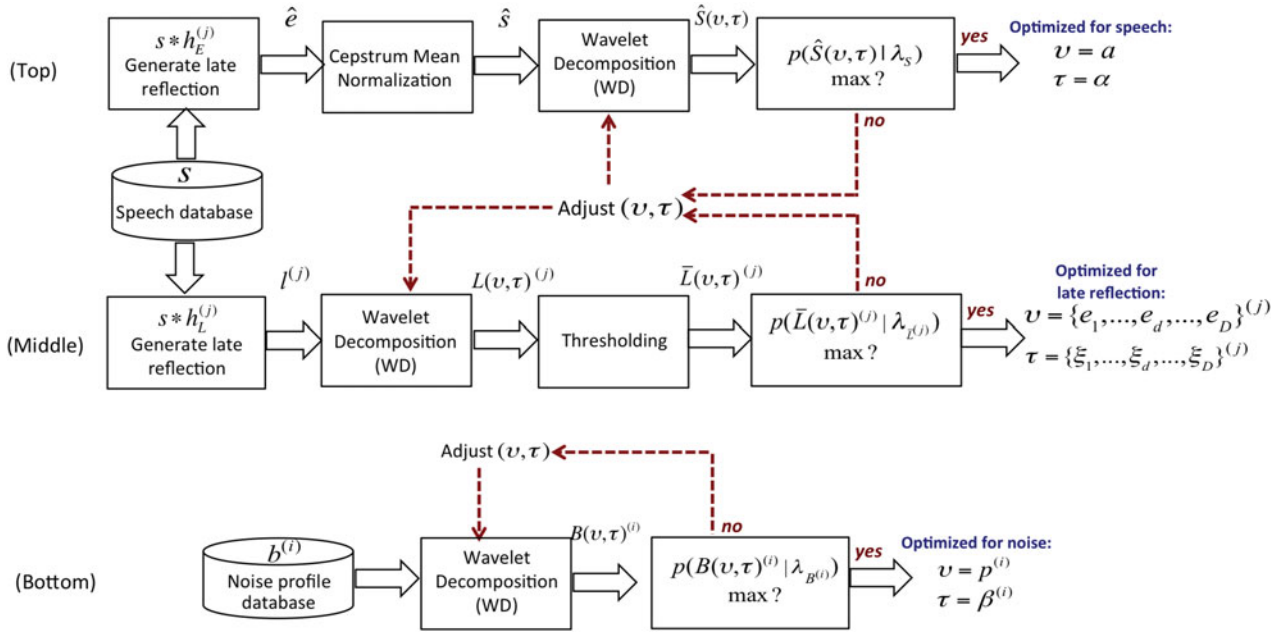

Fig. 2. Wavelet parameter optimization scheme for speech, late reflection and background noise.

ADDITIVE BACKGROUND NOISE

The same procedure is applied to the additive background noise as shown in Fig. 2: (Bottom), except for the creation of multiple noise profiles (i), representing different types of background noise. After decomposing the time domain signal $b^{(i)}$ through wavelet decomposition to $B(\upsilon,\tau)^{(i)}$, likelihood scores are computed using the corresponding noise model $\lambda_{B^{(i)}}$ (same model structure as that of λ_S). This model is trained using a noise database. The corresponding υ=$p^{(i)}$ and τ=$\beta^{(i)}$ that maximize the likelihood score are stored and associated with the corresponding noise profile.

The noise database is originally composed of different background noise recordings referred to as base noise (i.e. Mall, Hall, Crowd, Office, Vacuum, and Computer). To generalize to a variety of noise characteristics, additional entries are made by combining different types of the base noise. First, a simple piece-wise combination is performed and the resulting noise combination is further combined in the next level. To remove redundancy and suppress the increase of the entries, we measure the correlation of the resulting combinations and select the ones that are less correlated with existing noise entries. Thus, the expanded noise database referred to as noise profiles will provide more degree of freedom in characterizing various noise distributions. More detailed explanation regarding noise profiles is found in [12].

LATE REFLECTION

In the case of the late reflection, wavelet parameter tuning is shown in Fig. 2: (Middle). The late reflection $l^{(j)}$ for the corresponding reverberation time j (i.e. $T_{60}^{(j)}$) is generated using the clean speech database and the predetermined late reflection coefficients $h_l^{(j)}$ of the RIR. The late reflection boundary is predetermined experimentally as discussed in [6, 21]. Next, wavelet coefficients $L(\upsilon,\tau)^{(j)}$ are extracted through WD. In order to make $L(\upsilon,\tau)^{(j)}$ void of speech characteristics, thresholding is applied to $L(\upsilon,\tau)^{(j)}$. Speech energy is characterized with high coefficient values [17, 22] and thresholding sets these coefficients to zero as,

$$\bar{L}(\upsilon,\tau)^{(j)} = \begin{cases} 0, & |\, L(\upsilon,\tau)^{(j)} \,| > \text{thr}, \\ L(\upsilon,\tau)^{(j)}, & |\, L(\upsilon,\tau)^{(j)} \,| \leq \text{thr}, \end{cases} \tag{25}$$

thr is calculated similar to that in [22]. The thresholded signal $\bar{L}(\upsilon,\tau)^{(j)}$ is evaluated against a late reflection model $\lambda_{\bar{L}^{(j)}}$. D templates for every reverberation time $T_{60}^{(j)}$ are to be optimized for both scale $(\upsilon_1,\ldots\upsilon_D)^{(j)}$ and shift $(\tau_1,\ldots,\tau_D)^{(j)}$. These correspond to D preceding frames (refer to equation (4)) that cause smearing to the current frame of interest. We note that the effect of smearing is not constant, thus D templates are created and experimentally identified. The parameters υ and τ are adjusted and the corresponding υ=$\{e_1,\ldots,e_D\}^{(j)}$ and τ=$\{\xi_1,\ldots,\xi_D\}^{(j)}$ that result in the highest likelihood score are selected. We note that $\lambda_{\bar{L}^{(j)}}$ is trained using the synthetically generated late reflection data (during training) with thresholding applied.

C) Optimized wavelet-domain WF

The general expression of the conventional Wiener gain (un-optimized) at band m is expressed as

$$\kappa_m = \frac{S(\upsilon,\tau)_m^2}{S(\upsilon,\tau)_m^2 + L(\upsilon,\tau)_m^2 + B(\upsilon,\tau)_m^2},$$

where $S(\upsilon,\tau)_m^2$, $L(\upsilon,\tau)_m^2$, and $B(\upsilon,\tau)_m^2$ are the wavelet power estimates for the clean speech, late reflection, and background noise, respectively. Using the optimized values for υ and τ as described in Section III-B, we can compute the respective power estimates directly from the observed

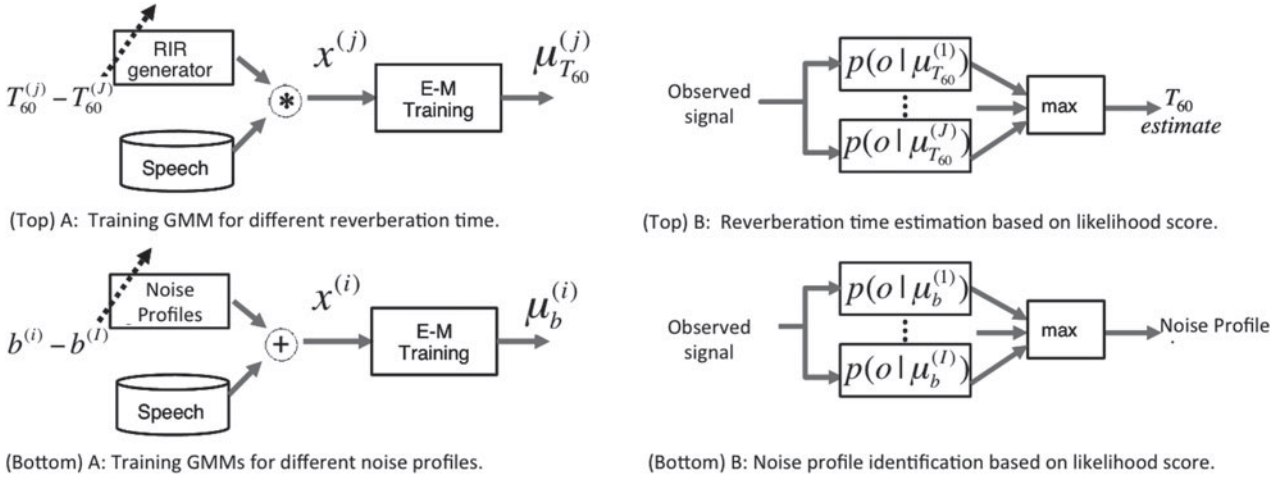

Fig. 3. Noise profile and reverberation time identification.

contaminated signal $O(\upsilon,\tau)_m$. Thus, the speech power estimate becomes

$$S(\upsilon,\tau)_m^2 \approx O(a,\alpha)_m^2, \tag{26}$$

the background noise power estimate $B(\upsilon,\tau)_m^2$ as

$$B(\upsilon,\tau)_m^2 \approx O(p^{(i)},\beta^{(i)})_m^2, \tag{27}$$

and the late reflection estimate $L(\upsilon,\tau)_m^2$ as

$$L(e_d^{(j)},\xi_d^{(j)})_m^2 \approx \begin{cases} O(e_1^{(j)},\xi_1^{(j)})_m^2, & d=1 \\ \frac{\sum_{k=1}^{d-1} O(e_k^{(j)},\xi_k^{(j)})_m^2}{d-1}+ & \\ O(e_d^{(j)},\xi_d^{(j)})_m^2, & \text{otherwise,} \end{cases} \tag{28}$$

WF is conducted by weighting the contaminated wavelet coefficient $O(\upsilon,\tau)_m$ with the Wiener gain as

$$O(\upsilon,\tau)_m(\text{enhanced}) = \mathrm{O}(\upsilon,\tau)_{\mathrm{m}}.\kappa_{\mathrm{m}}. \tag{29}$$

In equation (29), the Wiener weight κ_m dictates the degree of suppression of the contaminant to the observed signal at band m. If the contaminant power estimate is greater than the estimate of the speech power, then κ_m for that band may be set to zero or a small value. This attenuates the effect of contamination. On the other hand, if the power of the clean speech estimate is greater, the Wiener gain will emphasize its effect. Equation (29) is further processed with CMN prior to input to the ASR system.

IV. IDENTIFYING NOISE PROFILE AND REVERBERATION TIME

In Fig. 3 (Top A) we show the training of GMMs μ. Each GMM is composed of 256 mixture components. In this work, we experimentally set the step size of T_{60} to 20 ms. covering from 100 to 600 ms (each step size corresponds as a discrete entry (j). We use the RIR generator in [6] to synthetically create reverberant data $x^{(j)}$. In the case of the noise profiles in (Bottom A), we generated different background noise entry $x^{(i)}$ through synthetic superimposition using the speech database and the noise profiles. GMM architecture and training mechanism is the same as that used in the reverberant GMMs where $\mu_b^{(i)}$ GMMs are trained for each noise profile i.

The reverberation time (j) has a corresponding optimized wavelet parameters $\{e_1,\ldots,e_D\}^{(j)}$ and $\{\xi_1,\ldots,\xi_D\}^{(j)}$ while the noise profile (i) has $(p^{(i)},\beta^{(i)})$. During ASR use, it is necessary to identify the profile that corrupts the speech signal to retrieve the appropriate parameters. To identify reverberation time (j), a GMM-based classifier is employed in Fig. 3 (Top B) using the pre-trained reverberant models $\mu_{T_{60}}^{(j)}$. Subsequently, the profile (j) that leads to the best likelihood is selected. The same procedure is applied to the identification of noise profile (i) shown in (Bottom B).

V. EXPERIMENTAL EVALUATIONS

A) Experimental setup

TRAINING

We evaluate the proposed method in large vocabulary continuous speech recognition (LVCSR). The training database is the Japanese Newspaper Article Sentence (JNAS) corpus with a total of approximately 60 h of speech. The test set is composed of 200 sentences uttered by 50 speakers. The vocabulary size is 20 K and the language model is a standard word trigram model. Speech is processed using 25 ms-frame with 10 ms shift. The features used are 12-order MFCCs, ΔMFCCs, and ΔPower. The AM is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total.

TESTING

For testing, we used seven types of noise in the NAIST database [23]: mall, hall, crowd, office, vacuum cleaner, and computer noise. As a result of combining different types of

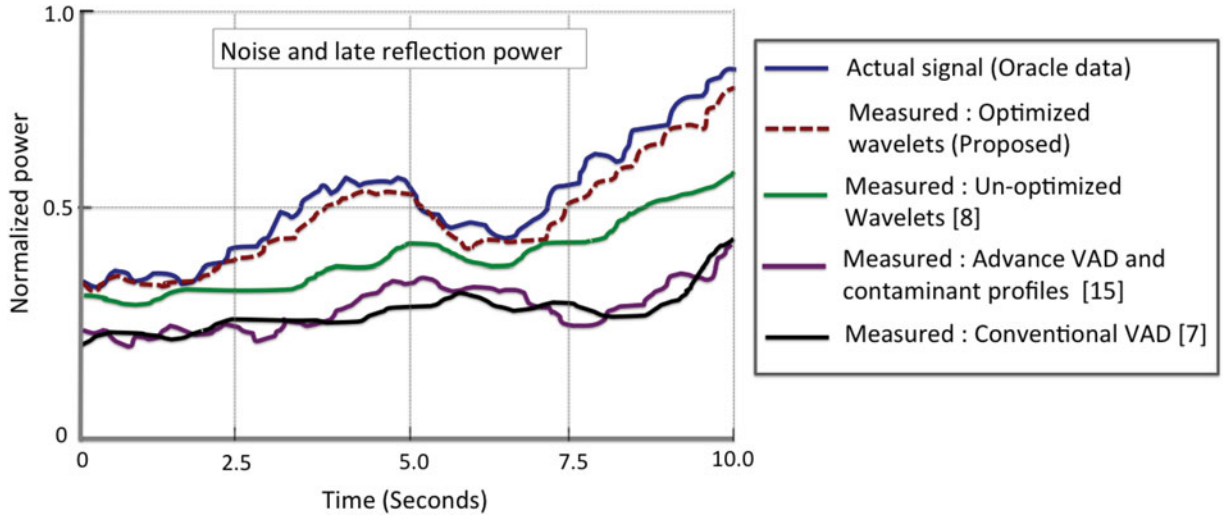

Fig. 4. Combined noise and late reflection power tracking.

noise from the noise database, 20 noise profiles are generated. We considered reverberation time T_{60} of 400 and 600 ms. with SNR of 20, 10, and 0 dB.

WAVELET FAMILY

In our work, we considered several wavelet families as shown in [24, 25]. Based on our experiments, the Daubechies wavelet was selected to better represent the speech signal. In the case of different noise types, the Symlet, Coiflet, and Meyer wavelets were selected in most cases. Lastly, the Gaussian derivative wavelets was selected for the late reflection. Wavelet selection is based primarily on the likelihood score between the actual signal and the AM, respectively. Hence, a noise type recorded at a particular environment condition may have a unique wavelet.

B) Noise and late reflection tracking performance

The advantage of optimizing the wavelets in estimating the signal of interest is shown in Fig. 4. In this experiment, noise and late reflection were super-imposed to the speech to synthesize the model in equation (5). The contaminant power was variably adjusted along the time axis to recreate a varying effect of contamination (i.e. amplitude). For simplicity, only the amplitude variability is shown. We note that for the oracle, we used the actual contaminant data for the power measurement while the observed signal (composite) is used for others. In this graph, the power envelope estimated using the proposed optimized wavelet parameter closely tracks the actual power of the contaminant. Obviously, the proposed wavelet optimization outperforms the other power estimation techniques.

C) GMM classification performance

The identification of the noise profile (i) and reverberation time (j) during recognition is vital in selecting the optimal wavelets. The overall performance of the proposed method depends on the accurate identification of these two. Given an observed reverberant and noisy data, we show in Table 1 the classification rate of the GMM classifier as a function of Gaussian mixtures. We have dyadically increased the mixture size in each training from 2 up to 512. With a smaller mixture, the classifier is unable to discriminate the the inherent characteristics of noise profiles and T_{60}, resulting to poor identification rate. As the mixture size is increased, the identification rate improves and saturates at 256 mixtures. This means that with a sufficient mixtures, the classifier is capable of resolving signal characteristics. We have found out that the identification of the noise profile and the reverberation time as discussed in Section IV works well even with only a few frames of data.

Table 1. GMM classification performance.

No. of mixtures (mix)	Noise profile (%)	Reverberation time (%)
2	5.0	10
4	14	22
8	26	35
16	38	42
32	47	55
64	62	75
128	85	94
256	**93**	**98**
512	94	98

D) Comparing ASR performance with other methods

We compare the proposed method against the following methods:

- **A)** No processing: Reverberant and noisy data processed with reverberant and noisy model;

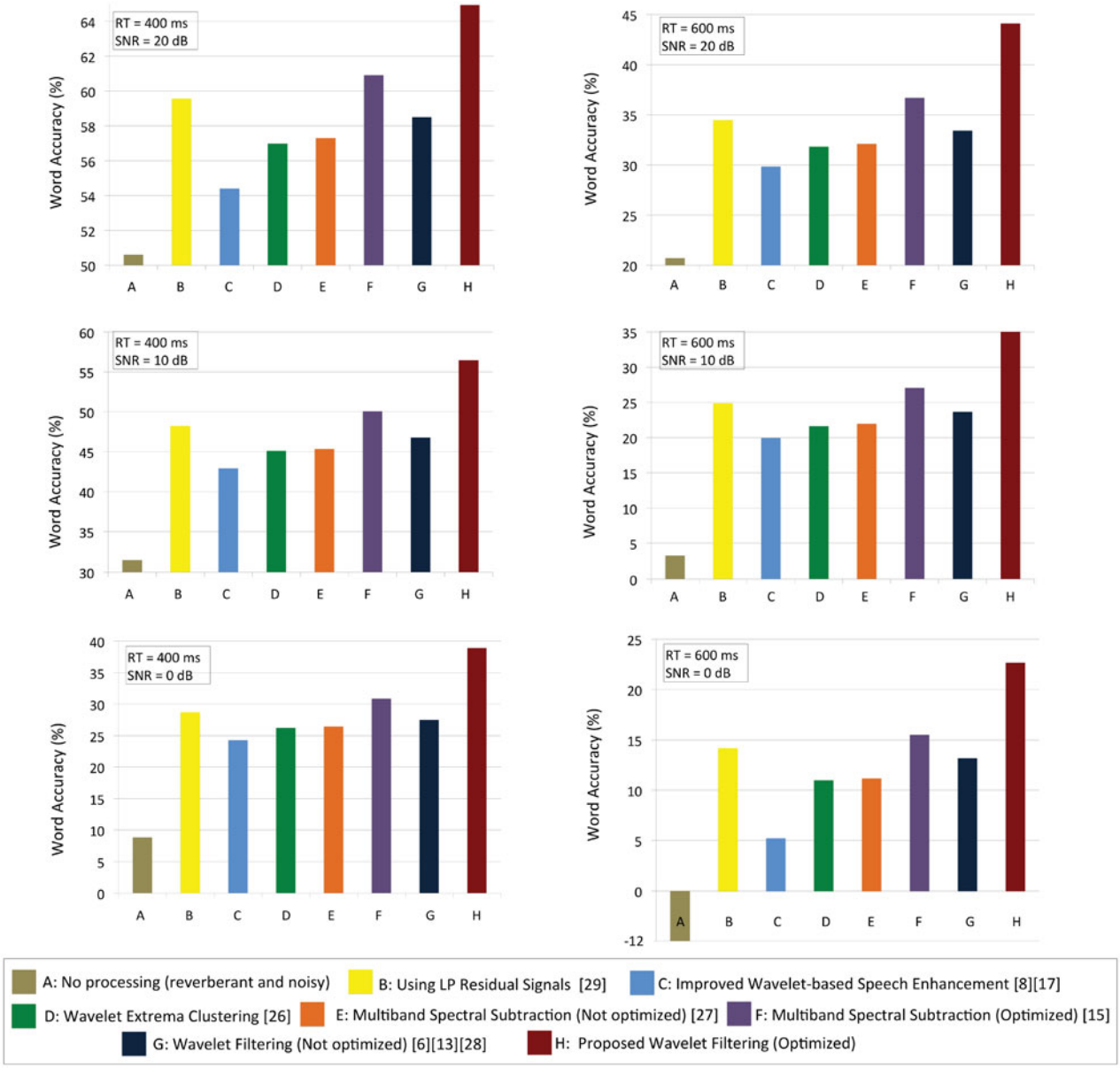

Fig. 5. ASR performance in word accuracy (averaged over all types of noise : Mall, Hall, Crowd, Office, Vacuum cleaner, and Computer noise.

- **B)** Based on LP residual signals: Reference technique based on Linear Prediction-based (LP) dereverberation analysis [26];
- **C)** Improved wavelet-based speech enhancement: Technique that incorporates VAD and contaminant-specific profiles for improved contaminant power estimation and improved performance [9, 17];
- **D)** Wavelet extrema clustering: Technique based on linear predictive coding (LPC) in the wavelet domain [27];
- **E)** Multi-band SS: Technique that employs suppression of the late reflection; optimization criterion is based on MMSE [21];
- **F)** Multi-band SS (ASR-optimized): Technique that suppresses the late reflection and at the same time maximizes the AM likelihood [6];
- **G)** Wavelet filtering (un-optimized): Technique that employs WF in the wavelet domain [7, 14, 28];
- **H)** Optimized wavelet filtering (Proposed method): Wavelet filtering is optimized using the AM.

While (C), (D), and (G) are wavelet-based techniques, (B), (E), and (F) are implemented in domains other than the wavelet. For the methods in (B)–(G), processing using the ETSI advanced front-end (AFE) [29] is applied to mitigate the effects of background noise.

The summary of the recognition performance comparing the proposed method and other existing methods described is shown in Fig. 5. In this figure, we show the word accuracy for reverberation time T_{60} = 400 and 600 ms. Each of the reverberant condition is also corrupted with background noise with SNRs 0 , 10, and 20 dB, respectively.

Table 2. Performance in word accuracy (%) attributed to the series of optimization.

	400 ms			600 ms		
	20 dB (%)	10 dB (%)	0 dB (%)	20 dB (%)	10 dB (%)	0 dB (%)
Mall noise						
(G) Wavelet filtering (un-optimized)	47.1	38.3	17.4	26.3	11.2	8.0
(H) Proposed wavelet filtering:						
Likelihood criterion (Section III-A.1)	49.5	41.4	21.8	28.7	15.9	11.5
Likelihood ratio criterion (Section III-A.2)	51.2	44.9	23.7	31.6	18.2	13.8
Optimized wavelet family and parameter (Section III-B)	53.9	47.5	27.7	36.8	24.6	15.8
Hall noise						
(G) Wavelet filtering (un-optimized)	52.6	39.0	20.3	28.8	14.0	7.50
(H) Proposed wavelet filtering:						
Likelihood criterion (Section III-A.1)	54.9	42.1	24.2	31.6	19.2	11.7
Likelihood ratio criterion (Section III-A.2)	56.6	44.8	26.8	34.7	22.9	14.6
Optimized wavelet family and parameter (Section III-B)	58.6	48.9	31.6	38.4	25.9	17.4
Crowd noise						
(G) Wavelet filtering (un-optimized)	61.1	49.7	29.5	34.4	27.8	14.5
(H) Proposed wavelet filtering						
Likelihood criterion (Section III-A.1)	63.5	53.4	33.7	37.3	31.7	18.4
Likelihood ratio criterion (Section III-A.2)	65.2	56.5	35.9	39.9	34.2	20.1
Optimized wavelet family and parameter (Section III-B)	67.5	60.2	40.8	44.2	38.9	24.3
Office noise						
(G) Wavelet filtering (un-optimized)	58.7	44.1	27.6	31.6	26.5	8.7
(H) Proposed wavelet filtering:						
Likelihood criterion (Section III-A.1)	61.8	47.6	31.3	36.8	29.2	13.9
Likelihood ratio criterion (Section III-A.2)	63.7	50.3	34.8	40.5	32.3	14.1
Optimized wavelet family and parameter (Section III-B)	65.1	54.2	38.9	49.4	35.7	19.2
Vacuum cleaner noise						
(G) Wavelet filtering (un-optimized)	63.9	53.1	32.9	37.6	30.8	19.1
(H) Proposed wavelet filtering:						
Likelihood criterion (Section III-A.1)	65.7	55.9	36.2	40.5	34.3	23.4
Likelihood ratio criterion (Section III-A.2)	67.3	58.5	40.4	43.1	37.2	25.0
Optimized wavelet family and parameter (Section III-B)	70.4	62.6	44.9	48.1	41.7	28.7
Computer noise						
(G) Wavelet filtering (un-optimized)	67.6	56.4	37.3	41.8	31.5	21.3
(H) Proposed wavelet filtering:						
Likelihood criterion (Section III-A.1)	69.2	59.7	42.7	43.5	35.2	24.6
Likelihood ratio criterion (Section III-A.2)	71.6	61.7	45.3	46.6	38.1	26.8
Optimized wavelet family and parameter (Section III-B)	74.2	65.3	49.6	51.6	43.4	30.5

Recognition performance is averaged over all types of noise (i.e. Mall, Hall, Crowd, Office, Vacuum cleaner, and Computer noise).

The results in Fig. 5 show that the proposed method outperforms existing wavelet-based methods in (C), (D), and (G). The main difference between the proposed method and these methods is the nature in which the wavelets are employed. While the latter indiscriminately use the same generic wavelet to represent both the contaminant and the desired signal, the proposed method uses a wavelet suitable for each of the signal via optimization in Section III-B. This results to an improved correspondence between the wavelet and the signal of interest.

The SS methods in (E) and (F) are based on Fourier transform processing. For these methods, the same basis function is used to process all of the signals of interest, while the proposed method in (H) employs wavelets that are optimized for a particular signal via training in Section III-C. Moreover, in the methods (E) and (F) there is no mechanism of improving the discrimination property among signal subspaces. As a result, the proposed method outperforms methods (E) and (F). The LP-based dereverberation scheme (B) is also based on the Fourier transform principle and this method is not tuned to the ASR. Moreover, it is very sensitive to the effects of the background noise. We note that at the very severe condition ($T60 = 600$ ms and SNR=0 dB), word accuracy becomes negative due to a large number of insertion errors together with substitution errors for unprocessed speech.

E) Effectiveness of the proposed wavelet optimization

In Table 2, we show the detailed recognition performance in word accuracy, when optimizing the wavelet family based on the likelihood criterion (Section II-A.1) and the likelihood ratio criterion (Section III-A.2). The former only deals with the correspondence matching between the wavelet family and the signal of interest, while the latter includes a mechanism to improve the subspace

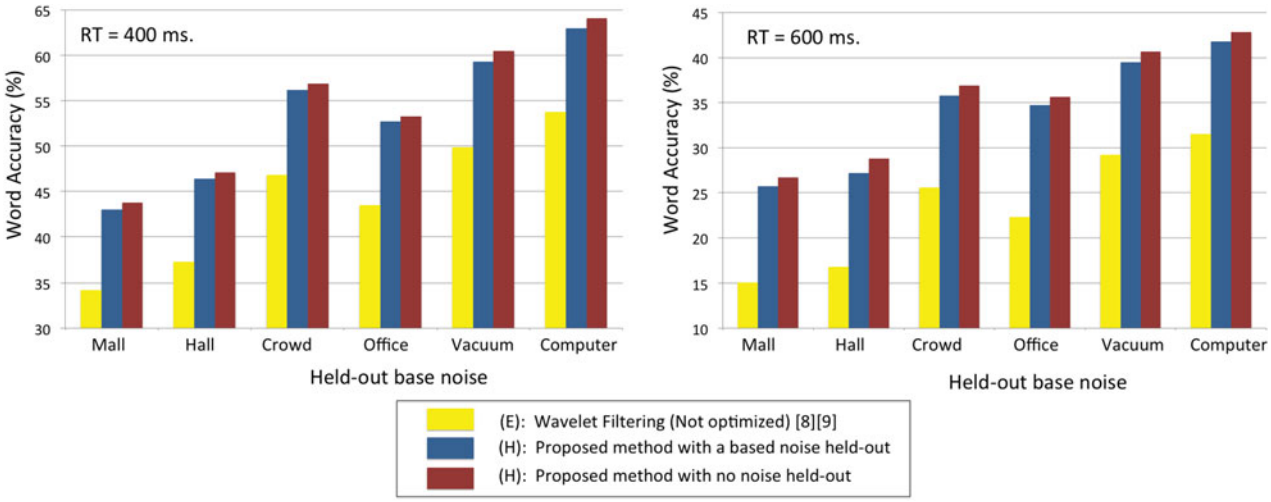

Fig. 6. Robustness to noise that are not enrolled in the profile database (averaged results of 20, 10, and 0 dB SNR).

discrimination among the signals of interest. In this table, we also show the effect of the proposed method (H) which includes both maximizing the likelihood ratio criterion (Section III-A.2) and optimizing the wavelet parameter (Section III-B). The existing wavelet-based filtering method (un-optimized) [7, 14] (G) is also provided. The effectiveness of the proposed method is confirmed in Table 2. For reference, both methods (C) and (G) use the Daubechies wavelet while method (D) uses a Quadratic spline wavelet.

F) Robustness to new noise types

The notion of expanding the original base noise into noise profiles is to find a representative of an unknown noise. We investigate the robustness of the proposed method in the event that a particular noise during testing is not covered in the noise profile database. To simulate this scenario, we held out the base noise together with its derivatives and compare its performance when it is not being held-out. The noise types that were excluded constitute as a new set of test data representing the new noise type. The comparative results are shown in Fig. 6. The difference in word accuracy between held-out (open test) and noise-enrolled, averaged over 20, 10, and 0 dB is negligible as shown in the figure, which means that the system is robust to noises that may not be present during enrolment. This may be attributed to the expansion of the noise database (i.e. noise profiles). The combination of different types of base noise as discussed in [12] has generated some noise profiles similar in characteristics to that of the held-out noise types. This renders the system to be robust. Moreover, the ability to select appropriate wavelet for different noise-types may have a positive impact as well.

VI. CONCLUSION

In this paper, we have presented the methods of optimizing the wavelet family and parameters using AM. The resulting optimized wavelets effectively estimate the power of the clean speech, late reflection, and background noise, respectively, from a contaminant signal. Thus, WF in the wavelet domain is improved. Since the optimization process is carried out using the AM, the enhanced speech signal is more likely to improve recognition performance. Currently, we deal with simple additive background noise since our method is primarily focused on dereverberation. In the future, we will further investigate a more sophisticated treatment of both dereverberation and denoising (combined); including the convolutive effect of noise. Further investigation in enhancing the current contaminant model to effectively address both reverberation and background noise will be a challenging task in our future work. Lastly, since the proposed method is a waveform enhancement technique, it can also be employed to train deep neural network-based ASR systems.

ACKNOWLEDGEMENTS

This work was mostly conducted while the first author was with Kyoto University and is a substantial extension to [12, 13].

REFERENCES

[1] Habets, E.: Single and multi-microphone speech dereverberation using spectral enhancement. *PhD Thesis*, June 2007.

[2] Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on ASSP, **27** (2), 1979, 113–120.

[3] Kim, W.; Kang, S.; Ko, H.: Spectral subtraction based on phonetic dependency and masking effects. Proc. IEEE Vis. Image Signal Process., **147**, 2000, 423–427.

[4] Lockwood, P.; Boudy, J.: Experiments with non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. Speech Commun., **11** (2–3) (1992), 215–228

[5] Soon, I.; Koh, S.; Yeo, C.: Selective magnitude subtraction for speech enhancement, in *Proc. The Fourth Int. Conf./Exhibition on High Performance Computing in The Asia Pacific Region*, 2000, vol. 2, 692–695.

[6] Gomez, R.; Kawahara, T.: Robust speech recognition based on dereverberation parameter optimization using Acoustic model likelihood. IEEE Trans. Audio, Speech and Lang. Proc., **18**, 2010, 1708–1716.

[7] Ambikairajah, E.; Tattersall, G.; Davis, A.: Wavelet transform-based speech enhancement, in *Proc. ICSLP*, 1998.

[8] Cohen, I.: Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process., **11**, 2003, 466–475.

[9] Ayat, S.; Manzuri-Shalmani, M.T.; Dianat, R.: An improved wavelet-based speech enhancement by using speech signal features. Comput. Electr. Eng., **32** (6), 2006, 411–425.

[10] Ayat, S.; Manzuri, M.; Dianat, R.; Kabudian, J.: An improved spectral subtraction speech enhancement system by using an adaptive spectral estimator, in *IEEE Canadian Conf. on Electrical and Computer Engineering*, 2005.

[11] Loizou, P.: Speech enhancement: theory and practice. *CRC Press*, Boca Raton, FL, 2007.

[12] Gomez, R.; Kawahara, T.: Denoising using optimized wavelet filtering for automatic speech recognition, in *Proc. of Interspeech*, 2011.

[13] Gomez, R.; Kawahara, T.: An improved wavelet-based dereverberation for robust automatic speech recognition, in *Proc. of Interspeech*, 2010.

[14] Gomez, R.; Even, J.; Saruwatari, H.; Shikano, K.: Distant-talking Robust speech recognition using late reflection components of room impulse response, in *ICASSP*, 2008.

[15] Gomez, R.; Kawahara, T.: Optimization of Dereverberation parameters based on likelihood of speech recognizer. in *Proc. of Interspeech*, 2009.

[16] Zelniker, G.; Taylor, F.: Advanced digital signal processing. *Marcel Dekker, Inc.*, New York, 1994.

[17] Sheikhzadeh, H.; Abutalebi, H.: An improved wavelet-based speech enhancement system, in *Proc. of Eurospeech*, 2001.

[18] Seltzer, M.: Speech-recognizer-based optimization for microphone array processing, in *Proc. of EEE Signal Processing Letters*, 2003.

[19] Kuttruff, H.: Room acoustics. *Spon Press*, London, 2000.

[20] Hirsch, H.-G.; Finster, H.: A new approach for the adaptation of HMMs to reverberation and background noise. Speech Commun., **50**, 2008, 244–263.

[21] Gomez, R.; Even, J.; Saruwatari, H.; Shikano, K.: Fast dereverberation for hands-free speech recognition, in *Proc. of the Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.

[22] Donoho, D.L.: Denoising by soft thresholding. *IEEE Trans. Inf. Theory,* **41**, 1995, 613–617.

[23] Yamade, S.; Matsunami, K.; Baba, A.; Lee, A.: Saruwatari, H.; Shikano, K.: Spectral subtraction in noisy environments applied to speaker adaptation based on HMM sufficient statistics, in *Proc. of ICSLP*, 2000.

[24] Daubechies, I.: Ten lectures on wavelets. *SIAM*, Philadelphia, PA, 1992.

[25] Misit, M.; Misiti, Y.; Oppenheim, G.; Poggi, J.: Wavelet toolbox user guide. *Mathworks*, Natick, MA, 2014.

[26] Yegnanarayana, B.; Satyaranyarana, P.: Enhancement of reverberant speech using LP residual signals. Proc. of IEEE Trans. on Audio, Speech and Lang. Proc., **8** (3), 2000, 267–281.

[27] Griebel, S.; Brandstein, M.: Wavelet transform extrema clustering for multi-channel speech dereverberation, in *IEEE Workshop on Acoustic Echo and Noise Control*, 1999.

[28] Gomez, R.; Kawahara, T.: Optimizing wavelet parameters for dereverberation in automatic speech recognition, in *Proc. of APSIPA*, 2010.

[29] Advanced Front-End Feature Extraction Algorithm, *ETSI Standard Document ES 202 050*, 2002.

Randy Gomez received M.Eng.Sci. in Electrical Engineering at the University of New South Wales (UNSW), Australia in 2002. He obtained his Ph.D. in 2006 from the Graduate School of Information Science, Nara Institute of Science and Technology (Shikano Laboratory), Japan. He was a Japan Society for Promotion of Science (JSPS) fellow at the Academic Center for Computing and Media Studies (Kawahara Laboratory), Kyoto University. Currently he is a senior scientist at Honda Research Institute Japan, Co. Ltd. His research interests include robust speech recognition, acoustic modeling and adaptation and multimodal interaction.

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including free large vocabulary continuous speech recognition software (http://julius.sourceforge.jp/) and the automatic transcription system for the Japanese Parliament (Diet). Dr. Kawahara received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. He was a general chair of IEEE Automatic Speech Recognition & Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He is an editorial board member of Elsevier Journal of Computer Speech and Language, APSIPA Transactions on Signal and Information, and IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is VP-Publications (BoG member) of APSIPA and a senior member of IEEE.

Kazuhiro Nakadai received a B.E. in electrical engineering in 1993, an M.E. in informationengineering in 1995, and a Ph.D. in electrical engineering in 2003, allfrom the University of Tokyo. He worked at Nippon Telegraph and Telephoneand NTT Comware Corporation from 1995 to 1999, and at the Kitano SymbioticSystems Project, ERATO, Japan Science and Technology Agency (JST) from 1999to 2003. He is currently a principal researcher for Honda ResearchInstitute Japan Co., Ltd., as well as having two visiting professorpositions, at Tokyo Institute of Technology and Waseda University from2011. His research interests include AI, robotics, signal processing,computational auditory scene analysis, multimodal integration and robotaudition. He is a member of the ASJ, RSJ, JSAI, IPSJ, HIS, and IEEE.