SPEECH ENHANCEMENT OPTIMIZATION BASED ON ACOUSTIC MODEL LIKELIHOOD FOR NOISY AND REVERBERANT ENVIRONMENT

Randy Gomez and Tatsuya Kawahara

Kyoto University, ACCMS,
Sakyo-ku, Kyoto 606-8501, JAPAN

ABSTRACT

Noise and channel contamination acoustically degrade the speech signal. To suppress the effects of degradation and recover the original signal, speech enhancement techniques are employed. In this paper, we focus on two simple and low-computational methods: Wiener filtering (WF) and spectral subtraction (SS). Conventionally, these are formulated with no relation with automatic speech recognition (ASR). We propose to optimize the conventional speech enhancement technique in relation with likelihood of the acoustic model. We also exploit these simple speech enhancement techniques that are originally designed for denoising, to address reverberation as well. In the experiment with real noisy and reverberant environments, we have achieved significant improvement in recognition performance using the proposed approach.

Index Terms— Robustness in ASR, Dereverberation, Denoising, Spectral Subtraction, Wiener Filtering

1. INTRODUCTION

Acoustic degradation is a common problem in speech recognition applications. There have been a lot of research involving speech enhancement that are specifically designed to suppress acoustic degradation of the speech signal caused by channel and noise. One of the widely used approaches is Wiener filtering (WF) [1] where short term estimates of the noise and speech are used in defining an adaptive filter to reduce as much noise energy while removing little speech energy as possible. A number of variants have been proposed and implementations in different domains such as time, frequency and wavelet [1] [2] are investigated. Another popular enhancement technique based on spectral subtraction (SS) [3] which removes the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal. A modification is given in [4] where multi-band is considered to deal with different effects of noise in different frequencies. Although these simple methods are widely used, they are formulated totally independent of the backend ASR systems.

Another approach which is linked with ASR or acoustic model likelihood is the feature transformation and adaptation [5] [6] [7]. Although these methods work well, they require a sufficient amount of adaptation data, and need some training to derive mapping parameters. These methods cannot be easily deployed in arbitrary environments especially when information of the room acoustics is not available.

In this paper, we focus on the simple enhancement algorithms: Wiener filtering (WF) and spectral subtraction (SS). Although there exist more sophisticated approaches, the enhancement schemes based on WF and SS are simple and fast to implement, which make its adoption to be effective in ASR applications. We first extend the original formulation of WF and SS to work in reverberant environments and then optimize the enhancement process in relation with ASR.

The paper is organized as follows; in Section 2, we show the method of extending both WF and SS to address reverberant conditions. In Section 3, we discuss the optimization of the scaling parameters used in WF and SS in the context of ASR followed by the RIR estimation in Section 4. Experimental conditions and results are given in Section 5, and we will conclude this paper in Section 6.

2. METHODS

The classical noisy speech model is given as,

$$y(n) = s(n) + d(n) \tag{1}$$

where $s(n)$ and $d(n)$ are the uncorrelated speech and noise signal respectively. To make use of the classical speech enhancements to work in reverberant scenario, we treat the reverberant signal analogous to that of Eq. (1). Thus, the reverberant model is given as,

$$x(n) = x_E(n) + x_L(n) \tag{2}$$

where $x_E(n)$ and $x_L(n)$ are the uncorrelated early and late reflections. The early reflections are composed of the direct signal and reflections in earlier time while the latter renders itself as coloration due to multiple reflections. In this paper, we consider both speech $s(n)$ and noise $d(n)$ are reverberant in nature. Assuming we can access the room impulse

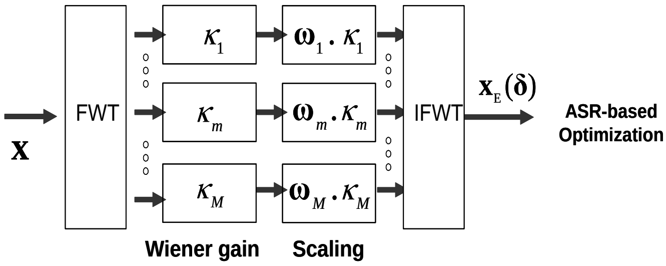

Fig. 1. Speech enhancement using Wiener filtering (WF)

response (RIR) $h(n) = [h_E(n)h_L(n)]$ and effectively identify its early and late components $h_E(n), h_L(n)$ [8][9] respectively, we can further rewrite Eq. (2) as,

$$x(n) = (s(n) + d(n)) * h_E(n) + (s(n) + d(n)) * h_L(n) \quad (3)$$

The power spectrum of the reverberant model in Eq. (2) can be estimated as:

$$|X(f)|^2 \approx |X_E(f)|^2 + |X_L(f)|^2 \quad (4)$$

where $X_E(f)$ and $X_L(f)$ are the magnitude spectra of the early and late reflections. By convention, we denote both $X_E(f)$ and $X_L(f)$ to contain both filtered speech and noise. Also, when referring to reverberant data $x(n)$, we assume a reverberant speech and reverberant noise as depicted in Eq. (3). In dealing with reverberation (both reverberant speech and noise), we are interested only in suppressing the effects of the late reflection since the early reflection is sensitive to microphone-speaker location. Moreover, the effect of early reflection is mostly mitigated with cepstral mean normalization (CMN) [8][9].

2.1. Wiener Filtering

The proposed Wiener filtering in the wavelet domain is a form of compression of the wavelet coefficients. By way of compression, the thresholding of the wavelet coefficients is avoided. The wavelet-based Wiener filtering [2] which is used in suppressing additive noise requires the calculation of Wiener gains given as,

$$\kappa_m = \frac{S(a)_m^2}{S(a)_m^2 + D(a)_m^2}, \quad (5)$$

where $S(a)_m^2$ and $D(a)_m^2$ are the speech and noise power respectively, calculated from the wavelet coefficients at scale a. Noise segments were detected using a voice activity detector (VAD). For the j^{th} contaminated wavelet coefficient in band m w_{mj}, the denoised wavelet coefficient is given as,

$$\tilde{w}_{mj}(denoised) = w_{mj}.\kappa_m, \quad (6)$$

The Wiener weighting κ_m dictates the degree of suppression of the contaminant to the observed signal. The enhanced wavelet coefficients are used to reconstruct the speech signal by inverse fast wavelet transform (IFWT).
This work of [2] is originally designed to suppress additive noise only. We expand it to deal with reverberant channel by suppressing the late reflections. Thus, the Wiener gain given in Eq. (5) is modified to,

$$\kappa_m = \frac{X_E(a)_m^2}{X_E(a)_m^2 + \delta_m X_L(a)_m^2}, \quad (7)$$

where $X_E(a)_m^2$ and $X_L(a)_m^2$ are the early and late reflection power respectively, calculated from the wavelet coefficients at scale a. Although $X_E(a)$ has relatively high power values than $X_L(a)$, the VAD method to select the correct segments may not be sufficient. Thus, a scaling parameter δ_m is introduced to minimize the error in calculating $X_E(a)_m^2$ and $X_L(a)_m^2$. We note that we can synthetically generate data using the clean speech and noise database together with the RIR [8][9]. Thus, we can calculate $\boldsymbol{\delta} = [\delta_1 ..., \delta_m, ..., \delta_M]$ that minimize the error between $\{X_E(a)_m^2, X_L(a)_m^2\}$ with the VAD and $\{X_E(a)_m^2, X_L(a)_m^2\}$ for the synthetically generated data. This process is similar to that in [8][9]. By applying the Wiener gains to the reverberant wavelet coefficients w_{mj} (analogous to Eq. 6), the enhanced wavelet coefficients are given as,

$$\tilde{w}_{mj}(enhanced) = w_{mj}.k_m. \quad (8)$$

The enhanced wavelet coefficients are converted back to the time domain through IFWT and we denote this as $x_E(\boldsymbol{\delta})$ to signify that only the early reflections are retained using $\boldsymbol{\delta}$. Fig. 1 illustrates the implementation of the modified WF where the reverberant and noisy speech signal is processed using a Fast Wavelet Transform (FWT). M subbands are created through FWT decomposition [11]. In our application we use five subbands to reflect that of [8][9]. Each of these subbands outputs a wavelet coefficient as a result of the fast wavelet structure. Then, the Wiener gains are calculated and the contaminated data is scaled by the Wiener gains. The early reflections (enhanced data) are then recovered through IFWT. Optimization

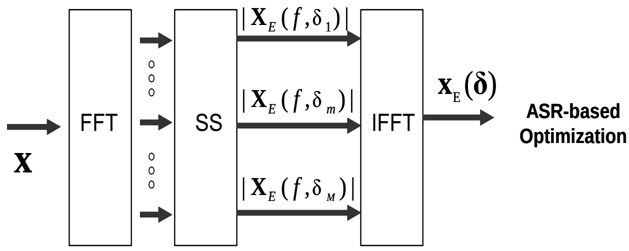

Fig. 2. Speech enhancement using spectral subtraction (SS)

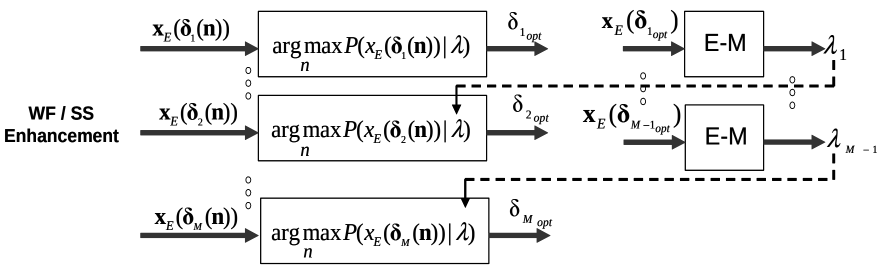

Fig. 3. ASR-based optimization of the scaling parameters.

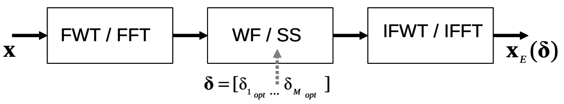

Fig. 4. Overall block diagram of the speech enhancement utilizing ASR-optimized scaling parameters.

of the scaling parameters based on ASR follows, which will be discussed in Section 3.

2.2. Spectral Subtraction

We will show the expansion of the conventional SS to address reverberation problems. As previously mentioned, we are interested in recovering only the early reflection and suppressing the late reflection. This can be done with multi-band SS [8][9]. Thus, the mth band power spectra of $X_E(f)$ is achieved through,

$$|X_E(f,\tau)| = \begin{cases} |X(f,\tau)|^2 - \delta_m |X_L(f,\tau)|^2 \\ \quad \text{if } |X(f,\tau)|^2 - \delta_m |X_L(f,\tau)|^2 > 0 \\ \beta |X_L(f,\tau)|^2 \quad otherwise \end{cases} \tag{9}$$

where β the flooring coefficient, $|X(f,\tau)|^2$ and $|X_L(f,\tau)|^2$ are the power spectra of the reverberant signal and power of the late reflection respectively, with a window period of τ. δ_m denotes the mth band scaling parameter. The multi-band scaling factors $\boldsymbol{\delta} = [\delta_1 ..., \delta_m, ..., \delta_M]$ are derived through an offline training which minimizes the error of the estimate $|X_L(f,\tau)|$ under the MMSE criterion. The values of $\boldsymbol{\delta}$ coefficients (through offline training), and the effective identification of the late components of the impulse response $h_L(n)$ are discussed in [8] [9]. Fig. 2 shows the block diagram of the SS implementation. First, the early reflection X_E are recovered as discussed in Eq. (9) and reverted back to $x_E(\boldsymbol{\delta})$ by IFFT.

Table 1. System specification used in evaluating the system

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	8000 Gaussian pdfs

3. OPTIMIZATION BASED ON ACOUSTIC LIKELIHOOD

In Section 2, the multi-band scaling parameters $\boldsymbol{\delta}$ are all set to initial MMSE-based values and in effect serve as a global weighting. In this section, we will discuss the optimization of $\boldsymbol{\delta}$, fine-tuning both WF and SS to be directly linked with ASR.

In Fig. 3, we show the ASR-based optimization of $\boldsymbol{\delta}$ where the scaling parameters in each band is sequentially optimized from band m=1 to m=M. The band coefficient to be optimized is allowed to change within a close neighborhood $n\triangle$ from its initial $MMSE$ value, where $n = \pm 1 ... N$ and $\triangle = 0.02$. The reverberant data $\boldsymbol{x}$ is enhanced using either multi-band WF/SS. Initially, we fix the rest of the scaling parameters to MMSE-based estimates except for the band to be optimized. Thus, for optimizing band $m = 1$, we generate $\boldsymbol{\delta_1}(\boldsymbol{n}) = [\,\delta_{1\ MMSE} + \boldsymbol{n}\triangle,\ \delta_{2\ MMSE},\ \delta_{m\ MMSE}, ...,\ \delta_{M\ MMSE}]$, and execute WF/SS using the generated coefficients. The resulting enhanced data $x_E(\boldsymbol{\delta_1}(\boldsymbol{n}))$ are evaluated using the HMM-based acoustic model which is trained with data processed with MMSE-based WF/SS parameters, denoted as $\lambda = \lambda_{MMSE}$. A likelihood score is computed for each of the data processed with different WF/SS conditions. Based on this result, $\delta(1)_{opt}$ that has the corresponding highest likelihood score is selected. Right after $\delta(1)_{opt}$ is found, the acoustic model is updated with data processed by WF/SS using $\delta(1)_{opt}$. The newly updated model λ_1 is then used in calculating the likelihood score for the next band

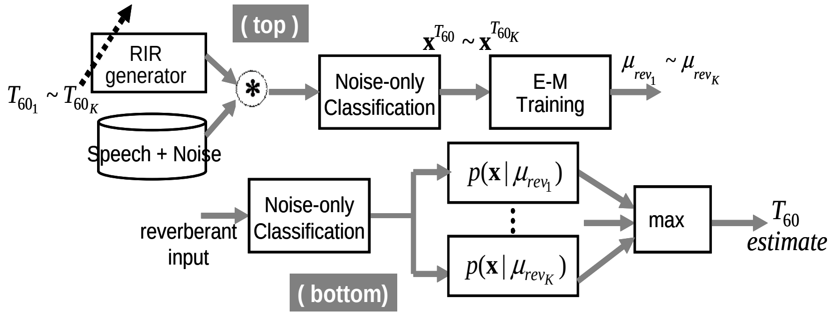

Fig. 5. Robust RIR Estimation.

and the process is repeated until the complete set of parameters $\delta_{1\,opt}$,...,$\delta_{M\,opt}$ are optimized. After the optimization, the reverberant data are processed with the proposed ASR-optimized WF/SS as shown in Fig. 4.

4. ROBUST RIR ESTIMATION

Since we need the RIR, we implement an automatic estimation of the RIR as opposed to physically measure it [8][9]. We have shown that due to the low resolution characterization of HMM to the speech signal compared to the RIR, rough estimate of the RIR is sufficient in HMM applications. The RIR can be modeled as having a decaying exponential energy,

$$h^2(n) \approx e^{(6\ \ln(10)/T_{60})\ l}, \tag{10}$$

where l is the discrete time sample, and T_{60} is the reverberation time. To effectively identify T_{60} in the presence of both convolutive speech and noise, we designed a GMM-based T_{60} classifier as shown in Fig. 5 (top). Reverberant speech and noise are synthetically generated $x^{T_{60_k}}$ with variable T_{60_k} to train GMMs μ_{rev_k}. To attain robustness, we employed the following; first, reverberant noise-only frames (occur in block segments during silence part of the clean speech) are used to train the GMM. This avoids the variability caused by the convolutive speech. From these reverberant noise-only block segments, we select only the frames that have low power to capture only the late reflection of the reverberant noise signal. We note that the late reflection renders itself as coloration in frequency due to multiple overlapping. This results in less sensitivity to noise types and SNR since noise information is smeared by the coloration effect. Finally, we use a larger mixture for the GMM (i.e. 256 mix). The use of a large number of mixture components makes the GMM sensitive to the higher resolution RIR. Fig. 5 (bottom) shows the actual identification of T_{60}. The reverberant speech and noise input is processed to classify noise-only frames. Then, likelihood is calculated given all of the GMMs with different T_{60_k}. The corresponding T_{60} that results in the highest likelihood score is selected and from this, the RIR is estimated using Eq. (10).

5. EXPERIMENTAL EVALUATION

5.1. Training and Testing Data

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set is composed of 200 utterances. Recognition experiments are carried out on the Japanese dictation task with 20K vocabulary. The language model is a standard word trigram model. The acoustic model is a triphone HMMs of 8000 Gaussian pdfs. A summary of the system specification is shown in Table 1.

We experimented using T_{60}=200 msec reverberation time. Reverberant training data are synthetically produced with the automatically generated RIR discussed in Section 4. The test data were recorded in a room with known reverberation time : T_{60}=200 msec. Thus, we used actual reverberant data for evaluation. Three types of noise are considered; office, vacuum cleaner, and white Gaussian noise. The signal-to-noise ratio (SNR) are 15 dB, 20 dB and 25 dB. The microphone-to-speaker distance is approximately 1.5 m. The noise source is also placed 1.5 m from the microphone with a 30 degrees angle relative to the microphone-to-speaker distance. In the experiments we use a total number of bands $M = 5$ which is consistent that of the former work [8][9].

5.2. Acoustic Model Training

We have shown the incremental optimization of the multi-band scale parameters in Section 3. This process selects the optimal scale factors $\boldsymbol{\delta}_{opt} = [\delta(1)_{opt}, ...\delta(m)_{opt}, ..., \delta(M)_{opt}]$. The acoustic model training is carried out as,

$$\lambda_{opt} = \arg\max_{\lambda_M} \prod_{r=1}^{R} P(\mathbf{x}_r^{\boldsymbol{\delta}_{opt}} | \mathbf{w}; \lambda_M),$$

where λ_{opt} is the desired acoustic model to be trained and later used by the ASR. λ_M is the Mth updated model which is the last model update in a series of model re-estimation as part of the optimization process discussed in Section 3. $\mathbf{x}_r^{\boldsymbol{\delta}_{opt}}$ is the enhanced utterance processed by WF/SS using the opti-

Table 2. Recognition Results in Word Accuracy

	office noise			vacuum cleaner noise			white gaussian noise		
Methods	15dB	20dB	25dB	15dB	20dB	25dB	15dB	20dB	25dB
$Testing$: Unprocessed $Training$: clean	23.4%	34.6%	40.3%	19.3%	32.2%	37.5%	25.6%	38.7%	42.0%
$Testing$: Unprocessed $Training$: Unprocessed	37.1%	43.5%	48.6%	35.4%	38.7%	42.6%	39.4%	45.1%	50.3%
$Testing$: SS $Training$: SS	51.8%	58.6%	63.2%	49.1%	57.3%	60.1%	52.8%	59.9%	64.7%
$Testing$: ASR-optimized SS $Training$: ASR-optimized SS	**61.4%**	**72.1%**	**75.9%**	**58.3%**	**70.1%**	**73.6%**	**63.4%**	**73.2%**	**77.1%**
$Testing$: WF $Training$: WF	52.3%	57.4%	61.8%	50.6%	56.4%	58.2%	53.6%	58.7%	62.9%
$Testing$: ASR-optimized WF $Training$: ASR-optimized WF	**62.5%**	**71.4%**	**74.1%**	**59.4%**	**68.3**%	**70.3%**	**64.7%**	**72.8%**	**76.5%**

mal scale parameters while $\mathbf{w}$ refers to its transcription. The training database has a total $r = R$ training utterances.

5.3. Recognition Performance

In Table 1, we show the recognition performance of the different methods. It is observed that enhancing the reverberant data using WF and SS is better than not processing the reverberant data at all. However, when WF and SS are optimized in relation with the ASR, further improvement in recognition performance is achieved. This is attributed to the fact that the ASR-optimized variants are capable of improving the model likelihood used by the ASR. The superior performance of the proposed method is consistent to all of the different SNRs and noise types in our experiment. We note that we test using real recording noisy and reverberant data.

6. CONCLUSION

We have extended two popular denoising techniques (WF and SS) to address reverberant speech and noise, and optimize each of these to be effectively used in ASR applications. Moreover, we have shown the process of embedding optimized enhancement in the acoustic model training. Improvement in performance is achieved as the enhancement procedure is closely linked to the improvement of the acoustic model likelihood. We have shown that the this concept works in both frequency and wavelet domain. In general, the optimization in relation to ASR is applicable to any speech enhancement algorithms and in any domains.

7. REFERENCES

[1] S. Vaseghi "Advanced Signal processing and Digital Noise reduction", Wiley and Teubner, 1996.

[2] E. Ambikairajah et. al., "Wavelet Transform-based Speech Enhancement" *In Proceedings International Conference on Speech and Language Processing ICSLP*, 1998

[3] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* pp 208-211, Apr. 1979.

[4] S. Kamath and P. Loizou,"A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* 2002.

[5] H. Hermansky et.al., "Data-driven Nonlinear Mapping for Feature Extraction in HMM" *ASRU Workshop*, 1999.

[6] T. Hwang et. al., "Feature Adaptation Using Deviation Vector for Robust Speech Recognition in Noisy Environment" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 1997.

[7] A. Torre et.al., "Non-linear Transformation of the Feature Space for Robust Speech Recognition" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2002.

[8] R. Gomez et.al. , "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[9] R. Gomez et.al., "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop HSCMA*, 2008

[10] R. Gomez, T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *Interspeech*, 2009.

[11] G. Strang, and T. Nguyen, "Wavelets and Filter Banks" *Wellesley-Cambridge Press*, 1996.