



Denoising Using Optimized Wavelet Filtering for Automatic Speech Recognition

Randy Gomez and Tatsuya Kawahara

Kyoto University, Academic Center for Computing and Media Studies (ACCMS), Sakyo-ku, Kyoto 606-8501, Japan

Abstract

We present an improved denoising method based on filtering of the noisy wavelet coefficients using a Wiener gain for automatic speech recognition (ASR). We optimize the wavelet parameters for speech and different noise profiles to achieve a better estimate of the Wiener gain for effective filtering. Moreover, we introduce a scaling parameter in the Wiener gain to minimize mismatch caused by distortion during the denoising process. Experimental results in large vocabulary continuous speech recognition (LVCSR) show that the proposed method is effective and robust to different noise conditions.

Index Terms: Speech recognition, Robustness, Denoising and Wavelet

1. Introduction

Background noise is often present in environments where automatic speech recognition (ASR) systems are deployed. A noisy signal results to degradation in recognition performance due to mismatch with the acoustic model (AM). Thus, speech processing techniques for noise suppression is one of the most important topics in ASR.

There are a number of denoising techniques, and most of them are based on the short term Fourier transform (STFT). In this paper, we focus on the wavelet transform because of its flexibility of using the analysis window of a variable length for different frequency bands. Moreover, we can manipulate its parameters to effectively discriminate the signal subspaces occupied between noise and speech [1]. Seminal works in wavelet denoising are based on waveshrink [2] and thresholding [3]. A more advanced method was proposed in [4]. This method introduces voice activity detection (VAD) and uses several threshold profiles for different types of noise. With the VAD, more accurate estimation of noise power is achieved. The use of noise profiles enables flexibility in switching to several thresholds to discriminate noise from speech.

Most of the existing wavelet methods [2][4] are generally designed to enhance the speech waveform, but this does not necessarily mean an improvement in ASR performance. Therefore, we propose an improved wavelet-based denoising method optimized for ASR. We optimize the wavelet parameters for speech and noise based on AM likelihood for improving the Wiener gain estimate. Wavelet filtering is performed by weighting the noisy wavelet coefficients with Wiener gains in multiple bands. This method was successfully applied to dereverberation in the previous work [5]. In this paper, we address its application to the denoising problem. Specifically in this application, two problems are addressed. First, there are a variety of noise types in real environments. Thus, we establish the notion of noise profiles to optimize specific wavelet parameters for each type

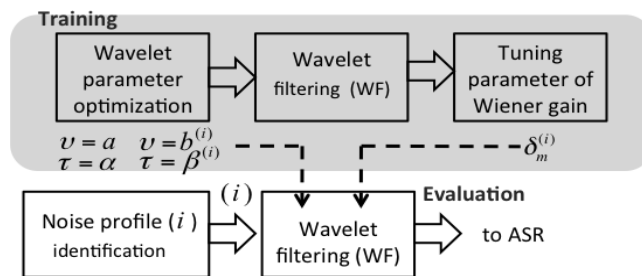


Figure 1: Block diagram of the proposed method.

of noise.

Second, even if a denoising method effectively suppresses noise, it often introduces distortion (i.e. residual noise) in the processed signal. The effects of distortion may be acceptable to human perception, but it may have a detrimental effect to ASR since it is another form of mismatch with the AM. One way of dealing with mismatch is to re-train the AM using the denoised data. However, there are many types of noise in real environments and it is impractical to re-train the AM for every noise condition. To deal with the residual noise, we introduce gain tuning in the Wiener gain. The parameter is optimized to minimize the mismatch between the denoised data (residual noise) and the noise data used in the AM training, and thus they will compensate the acoustic distortion caused by the wavelet filtering. During testing, an appropriate noise profile is identified and the corresponding optimized wavelet and tuning parameters for that profile are used to enhance the noisy speech input through the wavelet filtering prior to ASR. The whole process is depicted in Fig. 1.

The paper is organized as follows; Section 2 presents the proposed denoising method based on wavelet filtering by optimizing the wavelet parameters. In Section 3, we show the method of minimizing acoustic mismatch by tuning the Wiener gains. Then, noise profile identification is explained in Section 4. Experimental setup and ASR evaluation results are presented in Section 5. Finally, we conclude the paper in Section 6.

2. Wavelet Filtering for Denoising in ASR

2.1. Wavelet Parameter Optimization

A wavelet is generally expressed as

$$\Psi(v, \tau, t) = \frac{1}{\sqrt{v}} \Psi\left(\frac{t-\tau}{v}\right), \quad (1)$$

where t denotes time, v and τ are the scaling and shifting parameters respectively. $\Psi\left(\frac{t-\tau}{v}\right)$ is often referred to as the mother

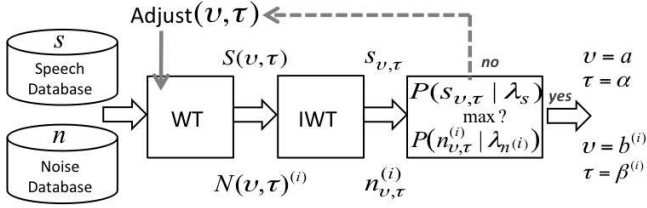


Figure 2: Wavelet parameter optimization scheme.

wavelet. Assuming that we deal with real-valued signal, the wavelet transform (WT) is defined as

$$F(v, \tau) = \int f(t) \Psi(v, \tau, t) dt, \quad (2)$$

where $F(v, \tau)$ is the wavelet coefficient and $f(t)$ is the time-domain function. With an appropriate training algorithm, we can optimize τ and v so that the wavelet captures specific characteristics of a certain signal of interest. The resulting wavelet is sensitive in detecting the presence of this signal given any arbitrary signal. In the wavelet filtering method, we are interested in detecting the power of clean speech and noise given a noisy observation. We optimize the wavelet parameters to detect clean speech and noise separately based on the AM likelihood as shown in Fig. 2. Since we are interested in speech subspace in general, optimizing a single wavelet to capture the general speech characteristics is sufficient. In the upper part of Fig. 2, we illustrate the optimization of the wavelet for clean speech. Wavelet coefficients $S(v, \tau)$, extracted through Eq. (2), are converted back to the time domain $s_{v, \tau}$ through inverse wavelet transform (IWT). Likelihood scores are computed using the clean speech acoustic model λ_s , a Gaussian Mixture Model (GMM) of 64 components. This is a text independent model which only captures the statistical information of the speech subspace. The process is iterated by adjusting v and τ . The corresponding $v=a$ and $\tau=\alpha$ that result to the highest score are selected.

The same procedure is applied to the case of noise, except for the creation of multiple profiles (i), representing different types of noise. Likelihood scores are computed using the corresponding noise model $\lambda_n^{(i)}$ (same model structure as that of λ_s). This model is trained using a noise database. The corresponding $v=b^{(i)}$ and $\tau=\beta^{(i)}$ that maximize the likelihood score are stored in the profile.

The noise database is originally composed of seven base noise, i.e. Car, Computer, Office, Crowd, Park, Mall and Vacuum cleaner. To generalize to a variety of noise characteristics, additional entries are made by combining different types of base noise. To remove redundancy and suppress the increase, we measure the correlation of the resulting combinations and select the ones that are less correlated with existing noises. Thus, the expanded noise database referred to as noise profiles will provide more degree of freedom in characterizing various noise distributions.

2.2. Wavelet Filtering

The general expression of the Wiener gain at window frame w and band m is expressed as

$$\kappa_{wm} = \frac{S(v, \tau)_{wm}^2}{S(v, \tau)_{wm}^2 + N(v, \tau)_{wm}^2}, \quad (3)$$

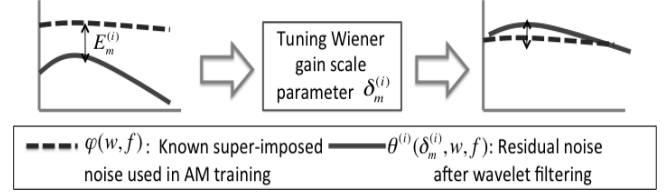


Figure 3: Tuning parameters of Wiener gain.

where $S(v, \tau)_{wm}^2$ and $N(v, \tau)_{wm}^2$ are wavelet power estimates for the clean speech and noise, respectively. And v and τ are the wavelet parameters, scale and shift. By using the optimized values for v and τ as discussed in Section 2.1, we can compute the speech and noise power estimates directly from the observed noisy signal $X(v, \tau)_{wm}$. Thus, the speech power estimate becomes

$$S(v, \tau)_{wm}^2 \approx X(a, \alpha)_{wm}^2, \quad (4)$$

and the noise power estimate $N(v, \tau)_{wm}^2$ as

$$N(v, \tau)_{wm}^2 \approx X(b^{(i)}, \beta^{(i)})_{wm}^2. \quad (5)$$

Wavelet filtering is conducted by weighting the noisy wavelet coefficient $X(v, \tau)_{wm}$ with the Wiener gain as,

$$X(v, \tau)_{wm}(\text{enhanced}) = X(v, \tau)_{wm} \cdot \kappa_{wm}. \quad (6)$$

In Eq. (6), the Wiener weight κ_{wm} dictates the degree of suppression of the contaminant noise to the observed signal at particular frame w and band m . If the noise power estimate is greater than the estimate of the speech power, then κ_{wm} for that band may be set to zero or a small value. This attenuates the effect of noise. On the other hand, if the power of the clean speech estimate is greater, the Wiener gain will emphasize its effect. The enhanced wavelet coefficients are converted back to the time domain through IWT and given to the ASR process.

3. Tuning Parameters of Wiener Gain

Denoising techniques often introduce distortion (i.e. residual noise), causing mismatch with the AM. To address this problem, super-imposition of a known noise was proposed [6][7]. Prior to training, a Gaussian noise is super-imposed to the clean speech database to train an AM [7]. Then, the same noise is super-imposed to the denoised speech during testing. However, it is not straightforward to determine the noise level super-imposed on the test data. Moreover, the method still depends on the noise types and denoising used. Thus, we introduce an additional scaling parameter δ in Eq. (3)

$$\kappa_{wm} = \frac{S(v, \tau)_{wm}^2}{S(v, \tau)_{wm}^2 + \delta_m^{(i)} N(v, \tau)_{wm}^2}, \quad (7)$$

to minimize the mismatch between the super-imposed noise (AM condition) and the residual noise. Tuning of $\delta_m^{(i)}$ is done offline and its concept is illustrated in Fig. 3.

We denote the spectrum of the known super-imposed noise as $\varphi(w, f)$ and the residual noise $\theta^{(i)}(\delta_m^{(i)}, w, f)$ for a given noise profile (i). Here, w and f are the frame index and frequency, respectively. $\theta^{(i)}(\delta_m^{(i)}, w, f)$ is derived by generating noisy data, i.e. adding noise to the speech database, and then denoising these noisy data with the wavelet filtering. Here only the frames where residual noise is dominant are used. The ar-

gument $\delta_m^{(i)}$ in $\theta^{(i)}(\delta_m^{(i)}, w, f)$ indicates that the residual noise spectrum is affected by the choice of $\delta_m^{(i)}$ through the wavelet filtering. The objective is to minimize the error E_m between the super-imposed noise $\varphi(w, f)$ and the residual noise $\theta^{(i)}(\delta_m^{(i)}, w, f)$ by adjusting $\delta_m^{(i)}$. For a given noise profile (i), the scaling parameter $\delta_m^{(i)}$ is optimized through minimum mean squared error (MMSE) criterion in each band m

$$E_m^{(i)} = \frac{1}{W} \sum_w \sum_{f \in B_m} |\varphi(w, f) - \theta^{(i)}(\delta_m^{(i)}, w, f)|^2, \quad (8)$$

where B_m is among the given set of bands. We used a total number of bands $M = 5$ similar to that in [5]. By this tuning of the Wiener gain, super-imposition of the known noise to the denoised utterance during testing is not needed anymore.

4. Noise Profile Identification

Each noise profile has corresponding optimized wavelet parameters ($b^{(i)}, \beta^{(i)}$ in Section 2.1) and tuning parameters of the Wiener gain ($\delta_m^{(i)}$ in Section 3). During testing in ASR, it is necessary to be able to classify the noise profile that corrupts the speech signal to retrieve the appropriate parameters and perform the proposed wavelet filtering. To identify the noise profile (i), a GMM-based classifier is employed. The GMMs ($\lambda_{n^{(i)}}$) are same as used in optimizing the wavelet parameters for the noise profiles discussed in Section 2.1. During testing, high-energy frames are removed from the input noisy speech and the remaining noise segments are evaluated with the GMMs. Subsequently, the profile (i) that leads to the best likelihood is selected. We have found out that the identification works well even with only a few frames of data.

5. Experimental Evaluations

We have evaluated 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 hours of speech. The test set is composed of 200 sentences uttered by 50 speakers. The vocabulary size is 20K and the language model is a standard word trigram model.

Speech is processed using 25 ms-frame with a 10 ms. shift. The features used are 12-order MFCCs, 12-order Δ MFCCs, and Δ Power. The AM is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. It is trained using the speech database with the super-imposition of Gaussian noise, that is different from those in the noise profiles. We note that in our proposed method, we use only a single AM in ASR for different noise and SNR conditions. We used seven types of real noise (base noise) in the NAIST database [7]: Car, Computer, Office, Crowd, Park, Mall and Vacuum cleaner. As the result of combination of the base noises, 20 noise profiles are derived.

In Tables 1, 2 and 3 we show the ASR performance in word accuracy for different methods in 20dB, 10dB and 0dB SNR. The accuracy in the clean condition is 93%. (A) is the result when the noisy test data is not processed, and recognized using an AM re-trained with the same noisy condition. (B) is the result of ETSI advanced front-end [8], a standard denoising for ASR. We also compare the performance with a denoising method based on Kalman filtering [9] in (C). In (D), we show the result of one of the best performing wavelet-based denois-

ing methods which employ VAD and noise profiles [4]. In (E), we show the performance of the conventional Wavelet filtering [1]. The proposed wavelet filtering method with wavelet parameter optimization is shown in (F). The optimization (Section 2.1) significantly improved the ASR performance, compared to the conventional wavelet filtering (E). The ASR performance is further improved by introducing the tuning parameter (Section 3) as shown in (G). The proposed method significantly outperforms both the conventional wavelet methods (D and E) and standard non-wavelet denoising methods (B and C).

Next, we investigated 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 some noise type and compare its performance when the noise is included in the noise profile database (i.e. (G)). The decrease in word accuracy (averaged over 20dB, 10dB and 0dB) shown in Fig. 4 between the two is very small, which means that the system is robust. The performance for the held-out noise condition is still better than that of the best performing denoising method by ETSI advanced front-end. The robustness of the system may be attributed to the expansion of the noise database (i.e. noise profiles) by combining different types of base noise. Note that the held-out noise type was not used to expand the noise profile database in this experiment.

6. Conclusion

We have presented an improved wavelet filtering, by optimizing the wavelet parameters to effectively estimate the power of the clean speech and the noise. This optimization is based on the AM likelihood, and results to a more accurate Wiener gain estimate for denoising. We have also proposed a method to compensate distortion caused by the wavelet filtering, by introducing a scale parameter in the Wiener gain. Since the tuning parameter is optimized to minimize the acoustic mismatch between the denoised data and the AM, ASR performance is also enhanced. In the future, we will further investigate the generalization of noise to be included in the noise profiles for more robustness to different noisy conditions.

7. References

- [1] E. Ambikairajah et. al., "Wavelet Transform-based Speech Enhancement" *In Proceedings of ICSLP*, 1998.
- [2] H.Y. Gao, "Wavelet Shrinkage Denoising", *In Proceedings of Computational Graphical Statistics* 1998.
- [3] D.L. Donoho, "Denoising by soft thresholding", *IEEE Trans. Info. Theory* 1995.
- [4] H. Sheikhzadeh and H. Abutalebi, "An Improved Wavelet-based Speech Enhancement System" *In Proceedings of Eurospeech*, 2001.
- [5] R. Gomez, T. Kawahara, "An Improved Wavelet-based Dereverberation for Robust Automatic Speech Recognition" *In Proceedings of Interspeech*, 2010.
- [6] D.V. Compernelle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System" *Computer Speech and Language* 1989.
- [7] S. Yamade, K. Matsunami, A. Baba, A. Lee, H. Saruwatari and K. Shikano, "Spectral subtraction in noisy environments applied to speaker adaptation based on HMM Sufficient Statistics", *In Proceedings of ICSLP*, 2000.
- [8] Advanced Front-End Feature Extraction Algorithm, *ETSI Standard Document ES 202 050*, 2002.
- [9] M. Fujimoto and Y. Akiri, "Noisy Speech Recognition using Noise Reduction Method based on Kalman Filter" *In Proceedings of ICASSP*, 2000.

Table 1: Evaluation results in word accuracy (20 dB SNR)

	Car	Computer	Office	Crowd	Park	Mall	Vacuum	average
(A) No processing	72.0%	69.3%	63.3%	64.8%	51.2%	43.0%	64.5%	61.2%
(B) ETSI [8]	87.3%	86.4%	78.4%	79.9%	62.5%	57.3%	81.2%	76.1%
(C) Kalman Filtering [9]	86.1%	85.2%	77.3%	78.5%	61.7%	56.9%	80.1%	75.1%
(D) Wavelet Denoising [4]	84.5%	83.6%	76.1%	76.4%	58.9%	55.2%	78.7%	73.4%
(E) Wavelet Filtering (WF) [1]	85.8%	84.3%	76.8%	77.8%	60.3%	55.7%	79.4%	74.3%
(F) Proposed WF	89.7%	88.3%	82.6%	83.5%	64.8%	59.0%	83.3%	78.7%
(G) Proposed WF + gain tuning	91.3%	89.2%	84.0%	84.7%	65.9%	62.6%	84.9%	80.3%

Table 2: Evaluation results in word accuracy (10 dB SNR)

	Car	Computer	Office	Crowd	Park	Mall	Vacuum	average
(A) No processing	59.2%	56.9%	47.6%	49.0%	38.5%	35.9%	53.7%	48.7%
(B) ETSI [8]	78.0%	75.8%	64.2%	65.6%	52.1%	50.5%	71.4%	65.4%
(C) Kalman Filtering [9]	77.1%	74.3%	63.4%	63.9%	50.8%	48.0%	70.2%	64.0%
(D) Wavelet Denoising [4]	72.7%	70.9%	61.2%	61.5%	47.6%	44.8%	68.3%	61.0%
(E) Wavelet Filtering (WF) [1]	73.4%	72.0%	62.1%	62.6%	48.4%	46.3%	69.0%	62.0%
(F) Proposed WF	82.8%	80.1%	68.7%	69.8%	56.4%	55.2%	75.4%	69.8%
(G) Proposed WF + gain tuning	84.6%	82.5%	71.4%	74.1%	58.9%	56.7%	77.0%	72.2%

Table 3: Evaluation results in word accuracy (0 dB SNR)

	Car	Computer	Office	Crowd	Park	Mall	Vacuum	average
(A) No processing	23.9%	20.1%	13.5%	15.3%	7.5%	5.3%	17.3%	14.7%
(B) ETSI [8]	48.3%	45.2%	31.6%	34.8%	24.7%	21.5%	29.1%	33.6%
(C) Kalman Filtering [9]	47.0%	43.7%	30.2%	33.4%	23.9%	20.7%	28.0%	32.4%
(D) Wavelet Denoising [4]	45.1%	41.4%	28.6%	31.8%	19.5%	18.9%	25.7%	30.1%
(E) Wavelet Filtering (WF) [1]	46.3%	42.1%	29.5%	32.6%	20.2%	19.1%	26.9%	31.0%
(F) Proposed WF	56.4%	54.3%	41.2%	42.5%	32.8%	28.7%	40.2%	42.3%
(G) Proposed WF + gain tuning	60.6%	58.7%	43.6%	45.8%	35.9%	33.0%	44.6%	46.0%

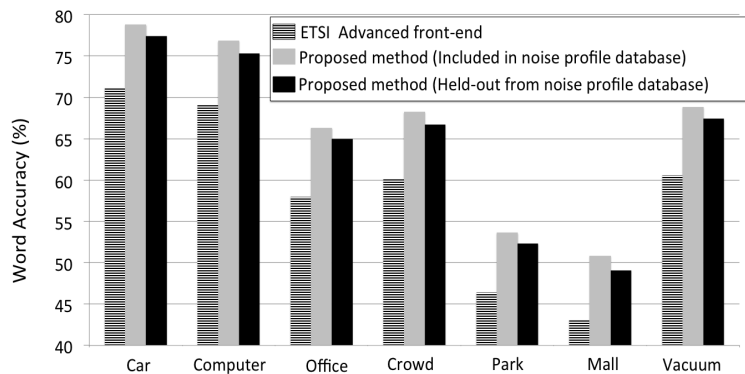


Figure 4: Robustness to noise that are not enrolled in the profile database (Averaged results of 20dB, 10dB and 0dB SNR).