Vol.2011-SLP-85 No.12
2011/2/5

ウエーブレットの最適化と雑音プロファイルを用いた雑音抑圧による頑健な音声認識

ゴメス・ランディ[†1]　河原　達也[†1]

本研究では、音声認識のためのウエーブレットに基づく雑音抑圧を雑音プロファイルと組み合わせることで改善を図る。学習時には、音声と種々の雑音プロファイル毎にウエーブレット変換のパラメータを最適化し、ウイナーフィルタのゲイン係数の推定の高精度化を図る。認識時には、雑音プロファイルを特定し、入力のウエーブレット係数を当該のウイナーゲインでフィルタリングする。さらに、ウイナーゲインにスケーリング係数を導入し、雑音抑圧に伴う歪みによるミスマッチを補償する。評価実験において、従来のウエーブレットに基づく手法と比較を行った。また、様々な雑音条件において頑健性の評価も行った。

Robust Speech Recognition Using Optimized Wavelet Denoising with Noise Profiles

Randy Gomez [†1] and Tatsuya Kawahara[†1]

In this paper, we improved the wavelet-based denoising method for automatic speech recognition (ASR) by using noise profiles. During training, we optimize the wavelet parameters for speech and different noise profiles to achieve a better estimate of the Wiener gain for effective filtering. Denoising is implemented by identifying the noise profile and filtering the noisy wavelet coefficients using a Wiener gain. In addition to wavelet filtering, we also introduce scale factors to the Wiener gain during decoding, to compensate for the mismatch caused by distortion during the denoising process. In our experimental evaluations, we compare our method with existing wavelet-based approach. We also conducted an experiment to test for robustness to different noise conditions.

1. Introduction

In real environment, automatic speech recognition (ASR) systems have to deal with background noise. When a speech signal is corrupted by noise, a mismatch with the acoustic model (AM) would result to degradation in recognition performance. 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 is proposed in[4][5]. 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 for improved discrimination between speech and noise subspaces.

Most of the existing wavelet methods[2][5] 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[6]. 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 in real environments. Thus, we establish the notion of the noise profiles to optimize specific wavelet parameters for each type 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

[†1] 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies (ACCMS), Kyoto University.
Randy Gomez is a research fellow of the Japan Society for Promotion of Science (JSPS).

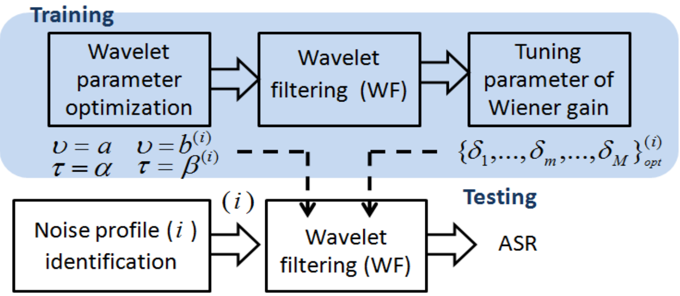
Training
Wavelet parameter optimization
Wavelet filtering (WF)
Tuning parameter of Wiener gain
$\upsilon = a$
$\tau = \alpha$
$\upsilon = b^{(i)}$
$\tau = \beta^{(i)}$
$\{\delta_1,\ldots,\delta_m,\ldots,\delta_M\}^{(i)}_{opt}$
Testing
Noise profile (i) identification
(i)
Wavelet filtering (WF)
ASR

図 1 Block diagram of the proposed method.

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 scale factors in the Wiener gain. The parameters are 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 improved 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.

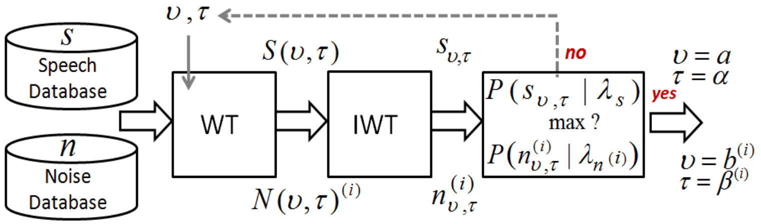

図 2 Optimization of the wavelet parameters through offline training.

2. Wavelet filtering for Denoising in ASR

2.1 Wavelet Parameter Optimization

A wavelet is generally expressed as

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

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

$$F(\upsilon,\tau) = \int f(t)\Psi(\upsilon,\tau,t)dt, \tag{2}$$

where $F(\upsilon,\tau)$ is the wavelet coefficient and $f(t)$ is the time-domain function. With an appropriate training algorithm, we can optimize τ and υ 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 acoustic model likelihood as shown in Fig. 2. Since we are only 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 $S(\upsilon,\tau)$, extracted through Eq. (2), are converted back to the time domain $s_{\upsilon,\tau}$. 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, adjusting υ and τ. The corresponding υ=a and τ=α 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 profile model $\lambda_{n^{(i)}}$ (same model structure as that of λ_s). This model is trained using noise-only frames from the noisy database. The corresponding noise wavelet profiles υ=$b^{(i)}$ and τ=$\beta^{(i)}$ are kept for evaluation.

The noise database is constructed by generating synthetic noise signals. By referring to the clean speech database, we can easily identify and segregate the noise-only frames to be used for training the noise models mentioned above. To generalize different noise characteristics, we increased the entry in the noise profiles by combining different types of base noise. Thus, the expanded 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 band m is expressed as

$$\kappa_m = \frac{S(\upsilon,\tau)_m^2}{S(\upsilon,\tau)_m^2 + \delta_m N(\upsilon,\tau)_m^2}, \tag{3}$$

where $S(\upsilon,\tau)_m^2$ and $N(\upsilon,\tau)_m^2$ are wavelet power estimates for the clean speech and noise, respectively. And υ and τ are the wavelet parameters scale and shift. δ_m is the tuning parameter, which will be discussed in Section 3. By using the optimized values for υ and τ as discussed in Section 2.1, we can compute the speech and noise power estimates directly from the observed noisy signal $X(\upsilon,\tau)$. Thus, the speech power estimate becomes

$$S(\upsilon,\tau)_m^2 \approx X(a,\alpha)_m^2, \tag{4}$$

and the noise power $N(\upsilon,\tau)_m^2$ estimate is given for frame-wise:

$$N(\upsilon,\tau)_m^2 \approx X(b^{(i)},\beta^{(i)})_m^2. \tag{5}$$

Wavelet filtering is conducted by weighting the noisy wavelet coefficient $X(\upsilon,\tau)$ with the Wiener gain as,

$$X(\upsilon,\tau)_m(enhanced) = X(\upsilon,\tau)_m \ . \ \kappa_m. \tag{6}$$

In Eq. (6), the Wiener weight κ_m dictates the degree of suppression of the contaminant noise to the observed signal. If the noise 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 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 inverse wavelet transform (IWT).

3. Tuning Parameters of Wiener Gain

Denoising techniques often introduce distortion (i.e. residual noise) causing mismatch to the AM. To address this problem, super-imposition of a known noise was proposed[7]. Prior to training, a known stationary noise is super-imposed to the clean speech database to train an AM. 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 additional scaling parameters in the Wiener gain to minimize the mismatch between the super-imposed noise (AM condition) and the residual noise (testing condition). This concept is illustrated in Fig. 3.

We denote the spectrum of the super-imposed noise as $\varphi(t,f)$ and the residual noise after the denoising as $\theta(\delta_m,t,f)$. Here, t and f are the frame index and frequency, respectively. The argument δ_m in $\theta(\delta_m,t,f)$ suggests that the residual noise spectrum is affected by the choice of δ_m through the wavelet filtering. The objective is to minimize the error E_m between the super-imposed noise $\varphi(t,f)$ and the residual noise $\theta(\delta_m,t,f)$ by manipulating δ_m. For a given noise profile (i), the scaling parameter δ_m is optimized through minimum mean squared error (MMSE) criterion in each band m

$$E_m = \frac{1}{T}\sum_t \sum_{f\in B_m} |\varphi(t,f) - \theta(\delta_m,t,f)|^2, \tag{7}$$

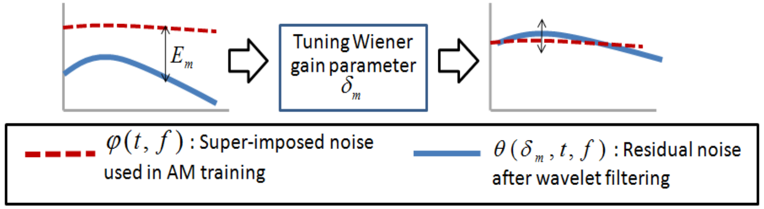
E_m
Tuning Wiener gain parameter δ_m
$\varphi(t, f)$: Super-imposed noise used in AM training
$\theta(\delta_m, t, f)$: Residual noise after wavelet filtering

図 3 Tuning parameters of Wiener gain.

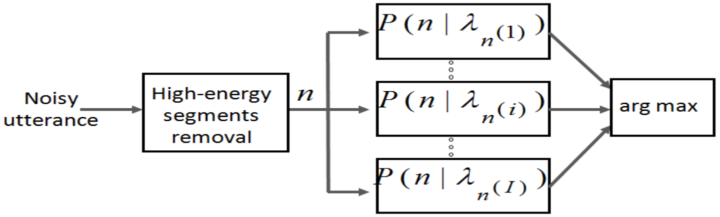
Noisy utterance
High-energy segments removal
n
$P(n \mid \lambda_{n(1)})$
$P(n \mid \lambda_{n(i)})$
$P(n \mid \lambda_{n(I)})$
arg max

図 4 Noise profile identification using Gaussian Mixture Model (GMM).

where B_m is among the given set of bands. We used a total number of bands $M = 5$[8)9)]. In this manner, $\delta_m^{(i)}$ is estimated for every noise profile i to be used during testing. By tuning 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 that corrupts the speech signal to retrieve the appropriate parameters and perform the improved wavelet filtering. A fast GMM-based classifier shown in Fig. 4 is employed in identifying the noise profile (i). After removing high-energy frames from the input speech, the remaining noise segments n are evaluated with the noise specific GMMs ($\lambda_{n(i)}$), which was explained in Section 2.1. 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 from 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 language model is a standard word trigram model. The acoustic model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. We used seven types of noises in the database[7)]: Car, Computer, Office, Crowd, Park, Mall and Vacuum cleaner. The noise profiles used in this experiment include the 21 combinations of the above mentioned noise.

In Tables 1 and 2, we show the ASR performance in word accuracy for different methods in 20dB and 10dB 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. In (B), we show the result of one of the best performing wavelet-based denoising methods which employ VAD and different noise statistical profiles[4)5)]. In (C), we show the performance of the conventional Wavelet filtering[1)]. The proposed wavelet filtering method with wavelet parameter optimization is shown in (D). The optimization (Section 2.1) significantly improved the ASR performance, compared to the conventional wavelet filtering in (C). The ASR performance is further improved by introducing the tuning parameters (Section 3) during decoding (E). The proposed method significantly outperforms the conventional 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. (E)). The decrease in word accuracy shown in Fig. 5 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 two reference methods[4)1)] (B) and (C) in Tables 1 and 2. The robustness of the system may be attributed to the expansion of the noise profile by combining different base noise. Note that the held-out noise type was not used to expand the noise profile database in this experiment.

	Car	Computer	Office	Crowd	Park	Mall	Vacuum	average
(A) No processing	72.0%	69.3%	63.3%	64.8%	51.2%	43.0%	62.5%	60.8%
(B) Wavelet Denoising[4]	85.8%	84.3%	77.8%	76.8%	60.3%	55.7%	79.4%	74.3%
(C) Wavelet Filtering (WF)[1]	84.5%	83.6%	76.4%	76.1%	58.9%	55.2%	78.7%	73.4%
(D) Proposed WF	**89.7%**	**88.3%**	**83.5%**	**82.6%**	**64.8%**	**59.0%**	**83.3%**	**78.7%**
(E) Proposed WF + gain tuning	**91.3%**	**89.2%**	**84.7%**	**84.0%**	**65.9%**	**62.6%**	**84.9%**	**80.3%**

表 1 Evaluation results in word accuracy (20 dB SNR)

	Car	Computer	Office	Crowd	Park	Mall	Vacuum	average
(A) No processing	59.2%	56.9%	47.6%	49.0%	28.8%	15.7%	41.6%	34.9%
(B) Wavelet Denoising[4]	73.4%	74.5%	63.6%	65.2%	36.1%	27.9%	72.8%	59.0%
(C) Wavelet Filtering (WF)[1]	72.7%	73.3%	62.2%	64.5%	35.3%	26.9%	71.4%	58.0%
(D) Proposed WF	**82.8%**	**80.1%**	**68.7%**	**69.8%**	**41.3%**	**33.2%**	**75.4%**	**64.4%**
(E) Proposed WF + gain tuning	**84.6%**	**82.5%**	**71.4%**	**74.1%**	**44.6%**	**35.9%**	**77.0%**	**67.5%**

表 2 Evaluation results in word accuracy (10 dB SNR)

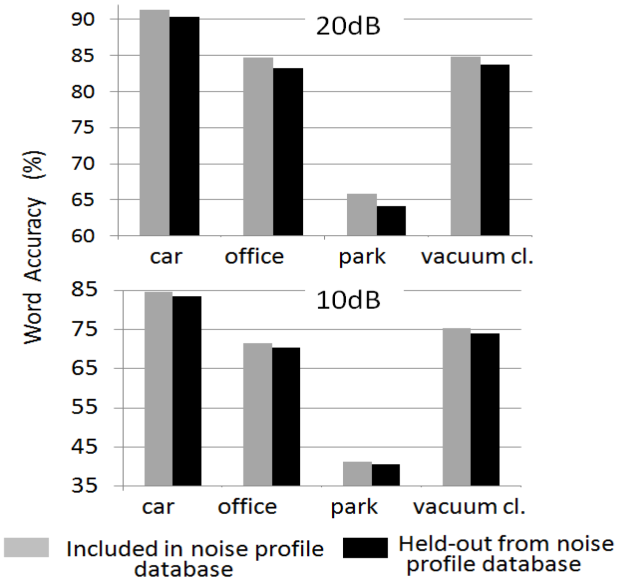

図 5 Robustness to noise that are not enrolled in the profile database.

6. Conclusion

We have proposed an improved wavelet filtering to address contaminant noise. This method optimizes the wavelet parameters to improve the Wiener gain estimate used in denoising. Moreover, the use of noise profiles enables the system to process different types of contaminant noise effectively. This results to a more accurate estimate of noise power for effective denoising. We have also introduced a mechanism to compensate distortion by the wavelet filtering, by tuning the Wiener gain during testing. Since the tuning parameters were optimized to minimize the acoustic mismatch between the denoised data and the AM, ASR performance is also enhanced. In the future, we will expand the formulation of this method to address both noise and reverberation problems.

参考文献

1) E. Ambikairajah et. al., "Wavelet Transform-based Speech Enhancement" *In Proceedings of the International Conference on Spokem Language Processing ICSLP*, 1998.

2) H.Y. Gao, "Wavelet Shrinkage Denoising", *In Proceedings of the Computational Graphical Statistics* 1998.

3) D.L. Donoho, "Denoising by soft thresholding", *In Proceedings of the IEEE Trans-*

action on Information Theory 1995.

4) H. Sheikhzadeh and H. Abutalebi, "An Improved Wavelet-based Speech Enhancement System" *In Proceedings of Eurospeech*, 2001.

5) S. Ayat, M.T. Manzuri-Shalmani and R. Dianat, "An Improved Wavelet-based Speech Enhancement by Using Speech Signal Features" *Elsevier*, 2006.

6) R. Gomez, T. Kawahara, "An Improved Wavelet-based Dereverberation for Robust Automatic Speech Recognition" *In Proceedings of Interspeech*, 2010.

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 the International Conference on Spokem Language Processing ICSLP*, 2000.

8) R. Gomez, J. Even, H. Saruwatari and K. Shikano, "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceedings of the IEEE International Conferences on Acoustics Speech and Signal Processing ICASSP*, 2008.

9) R. Gomez, J. Even, H. Saruwatari and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *In Proceedings of IEEE Workshop on Hands-free Speech Communication and Microphone Array HSCMA*, 2008.