PARAMETER SETTING OF NOISE REDUCTION FILTER USING SPEECH RECOGNITION SYSTEM

Tomomi Abe[1], Mitsuharu Matsumoto[2] and Shuji Hashimoto[1]
[1]*Waseda university, 55N-4F-10A, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan*
[2]*The University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo, 182-8585, Japan*

Keywords: Speech recognition system, Parameter optimization, Recognition-based approach, Nonlinear filter.

Abstract: This paper describes parameter setting of noise reduction filter using speech recognition system. Parameter setting problem is usually solved by maximization or minimization of some *objective* evaluation functions such as correlation and statistical independence. However, when we consider a single-channel noisy signal, it is difficult to employ such objective functions. It is also difficult to employ them when we consider impulsive noise because its duration is very small to use this assumption. To solve the problems, we directly use a speech recognition system as evaluation function for parameter setting. As an example, we employ time-frequency ε-filter and Julius as a filtering system and a speech recognition system, respectively. The experimental results show that the proposed approach has a potential to set the parameter in unknown environments.

1 INTRODUCTION

Although filtering plays an important role in acoustical signal processing, it is often necessary to determine some filter parameters to obtain the desired outputs. For instance, in utilizing comb filter (Lim et al., 1978), it is necessary to estimate the pitch of the speech signal as the parameter to reduce the noise signal adequately. ε-filter can reduce the acoustical noise while preserving speech signal if the parameter ε is adequately set (Harashima et al., 1982). To set the parameter adequately, many studies have been reported in the past. Mrazek et.al. proposed de-correlation criterion to select the optimal stopping time for nonlinear diffusion filtering (Mrazek and Navara, 2003). Sporring et.al. studied the behavior of generalized entropies, and employed it for determining the parameters (Sporring and Weickert, 1999). Umeda employs mutual information for blind deconvolution using independent component analysis (ICA) (Umeda, 2000). We also reported a parameter optimization algorithm based on signal-noise de-correlation (Matsumoto and Hashimoto, 2009; Abe et al., 2009).

However, it is sometimes difficult to set objective criterion for parameter setting practically. For instance, in the single-input single-output filtering system, we only have a single-channel noisy signal, that is, the original signal and noise are unknown. Although we usually set the assumption with regard to the relation between the signal and noise (e.g. decorrelation between signal and noise, or statistically independence between signal and noise), as we only have a single-channel noisy signal, it is not always easy to use these assumptions in single-channel filtering process. When we consider the impulsive noise such as clapping and percussion sounds, it becomes more difficult to employ the relation between signal and noise because its duration is very small to use this assumption. In other words, the signal-noise relation often could not be used. However, these types of noise often affects the recognition results seriously in spite of its instantaneous corruption. Moreover, the tuned filter output may not be always adequate for speech recognition system because filter and speech recognition system are separately tuned using different criterion.

To handle such cases, we pay attention to some *subjective* information in acoustical signal. For example, even when we cannot employ any objective assumptions, if we can generate the known sounds (words) from the system itself and use them as feedback, we can use the knowledge on the word to set the parameters. We can tune the parameter so that the filter output is the most target word-like by using speech recognition system. When we consider a robot system for simple tasks such as a cleaning robot, the number of the required words may be very small (for instance, from a few words to 10 words) com-

Abe T., Matsumoto M. and Hashimoto S..
PARAMETER SETTING OF NOISE REDUCTION FILTER USING SPEECH RECOGNITION SYSTEM.
DOI: 10.5220/0003054903870391
In *Proceedings of the International Conference on Fuzzy Computation and 2nd International Conference on Neural Computation* (ICNC-2010), pages 387-391
ISBN: 978-989-8425-32-4

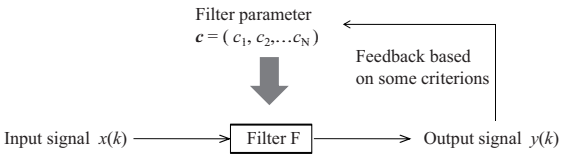


Figure 1: Parameter setting of filter system.

pared to that in the daily conversation. In these cases, it may be better that the filter is over-tuned for the objective words to make the speech recognition system robust for these words, even though the speech recognition system fails to recognize the other words. To evaluate our criterion, we employ time-frequency ε-filter (TF ε-filter) (Abe et al., 2007) and "Julius" (Lee et al., 2001; Lee and Kawahara, 2009; Kawahara et al., 2000) as a filtering system and a speech recognition system, respectively.

This paper is organized as follows. In Sec.2, we describe the concept of our approach and discuss the merits of the proposed approach compared to the other approaches. We also briefly explain the algorithm of Julius. In Sec.3, we explain TF ε-filter to clarify the problems. In Sec.4, we conduct some experiments and show the results to confirm the effectiveness of the proposed method. Discussions and conclusion are given in Sec.5.

2 PARAMETER SETTING USING SPEECH RECOGNITION SYSTEM

Let us start with a typical parameter setting problem to explain our approach clearly. Consider a filtering system for acoustical signal processing as shown in Fig.1. In Fig.1, $x(k)$ is a filter input. $y(k)$ is a filter output of filter F. We assume that the filter F has N parameters, which determine the filter characteristics. The parameter set is defined as $\mathbf{c} = (c_1, c_2, \cdots, c_N)$. Note that we do not mind the filtering types in Fig.1.

Parameter setting problem is usually solved as follows:

1. Definition of an Evaluation Function. At first, we set an evaluation function to evaluate whether the filter output is good or not. For instance, signal-noise decorrelation or statistically independence of signals are often used.

2. Maximization or Minimization of the Evaluation Function. After setting evaluation function, we set the filter parameters, and evaluate whether the output is adequate or not by using evaluation function. The optimal parameters are obtained as the parameters, which maximize or minimize the evaluation function.

Hence, it is considered that the main problem of parameter setting is how to set the evaluation function.

As an example, let us consider a noise reduction problem. The input signal includes not only the original speech signal but also noise in this case. In the filtering system, we only have a single-channel noisy signal, that is, the original speech signal and noise are unknown. Although we usually set the assumption with regard to the relation between the signal and noise (e.g. decorrelation between signal and noise, or statistically independence between signal and noise), as we only have a single-channel noisy signal, it is not always easy to use these criterions in single-channel filtering process. When we consider the impulsive noise such as clapping and percussion sounds, it becomes more difficult to employ this approach because its duration is very small to use this assumption. In other words, the signal-noise relation often could not be used. However, these types of noise often affects the recognition results seriously in spite of its instantaneous corruption.

To solve the problem, we directly utilize a speech recognition system to set the parameter of filter instead of the relation between signal and noise. The merits of this criterion are summarized as follows:

1. Simple Implementation. When we assume some relations between signal and noise such as decorrelation between signal and noise, or statistically independence between signal and noise, we need at least two signals to evaluate the relation. It is not always easy to employ the criterion when we only have a single-channel output. On the other hand, in our criterion, we only have to check whether the output signal is target speech or not. The checking process is simple, and does not require any other signals except the output.

2. Wide Application Range. Unlike objective evaluation function, this approach can be used regardless of noise types. Our approach can handle not only the noise occurring continuously like background noise but also the noise occurring infrequently like impulsive noise such as clapping and percussion sounds.

It is also considered that we do not need to care the recognition system itself. We can use any recognition system and regard it as an black box. We do not need to know the inside of the recognition system. We just ask the recognition system, and it react to us like

an artificial person who has a unique personality. The process is similar to the process of questionnaire to human.

Even when we set the system in the unknown space and can not set any objective assumptions, if we have sufficient sample sounds and can generate them there, we can set the filter system by tuning the parameter so that the filter output is the most generated word-like.

3. Affinity of Recognition System. The parameter is tuned so that the probability of the target speech in speech recognition system is the highest. Hence, the output may have some gaps from the objective perspective such as mean square error. However, when we employ the total system combining the filtering system and recognition system such as robot auditory, these types of gaps may give us some merits rather than demerit. Because the filter is tuned adequately not from objective perspective but from subjective perspective for recognition system itself. In this research, we use the open-source speech recognition software "Julius" (Lee et al., 2001; Lee and Kawahara, 2009; Kawahara et al., 2000) as a speech recognition system.

Let us define W, X, $p(W)$, $p(X)$ and $p(X|W)$ as the strings of the word, the input signal, the probability of W, the probability of X and the posterior probability of X for W, respectively. Julius decodes the word W from the input signal X by using the given acoustic model $p(X|W)$ and linguistic model $p(W)$. $p(W|X)$, the posterior probability of W for X, is calculated based on Bayes' theorem as follows:

$$p(W|X) = \frac{p(W) * p(X|W)}{p(X)}, \tag{1}$$

for the word W. In Eq.1, $p(X)$ is normalization factor which has no effect on determination of the word W and can be ignored. The estimated word $\hat{W}$ can be obtained as the word which maximizes the posterior probability $p(W|X)$. It can be described as follows:

$$\begin{aligned} \hat{W} &= \arg\max_W p(W|X) \\ &= \arg\max_W p(W) * p(X|W) \\ &= \arg\max_W \{\log\ p(W) + \log\ p(X|W)\}. \end{aligned} \tag{2}$$

We can assume that the linguistic model $p(W)$ is identical when we conducted a simple word recognition. The output of filter should become the most objective word-like from the acoustic perspective without linguistic model. In other words, the optimal parameter $\mathbf{c_{opt}}$ can be obtained as follows:

$$\mathbf{c_{opt}} = \arg\max_c \log\ p(X|W). \tag{3}$$

3 TIME-FREQUENCY ε-FILTER

In this section, we briefly describe the TF ε-filter algorithm. TF ε-filter is an improved ε-filter applied to the complex spectra along the time axis in time-frequency domain.

Let us define $x(k)$ as the input signal sampled at time k. In TF ε-filter, we firstly transform the input signal $x(k)$ to the complex amplitude $X(\kappa,\omega)$ by short term Fourier transformation (STFT). κ and ω represent the time frame in the time-frequency domain and the angular frequency, respectively. κ and ω are discrete numbers. Next we execute a TF ε-filter, which is an ε-filter applying to complex spectra along the time axis in the time-frequency domain. In this procedure, $Y(\kappa,\omega)$ is obtained as follows:

$$Y(\kappa,\omega) = \sum_{i=-Q}^{Q} \frac{1}{2Q+1} X'(\kappa+i,\omega), \tag{4}$$

where the window size of ε-filter is $2Q+1$,

$$\begin{aligned} &X'(\kappa+i,\omega) \\ &= \begin{cases} X(\kappa,\omega) & (||X(\kappa,\omega)| - |X(\kappa+i,\omega)|| > \varepsilon) \\ X(\kappa+i,\omega) & (||X(\kappa,\omega)| - |X(\kappa+i,\omega)|| \le \varepsilon), \end{cases} \end{aligned} \tag{5}$$

and ε is a constant. Then, we transform $Y(\kappa,\omega)$ to $y(k)$ by inverse STFT.

By utilizing TF ε-filter, we can reduce not only small amplitude stationary noise but also large amplitude nonstationary noise. It does not require the model not only of the signal but also of the noise in advance. It is easy to be designed and the calculation cost is small. Moreover, as it can reduce not only Gaussian noise but also impulsive noise such as clapping and percussion sounds, it is expected that it becomes a good example to show the effectiveness of our approach. See the reference (Abe et al., 2007) if the reader would like to know the details.

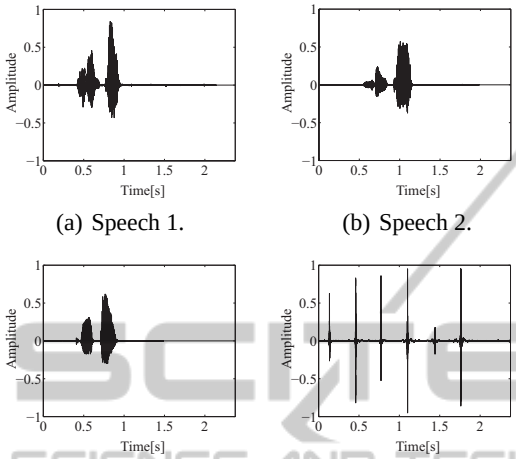
In TF ε-filter, ε is an essential parameter to reduce the noise appropriately (Abe et al., 2009). If ε is set at an excessively large value, the TF ε-filter becomes similar to linear filter and smooths not only the noise but also the signal. On the other hand, if ε is set to an excessively small value, it does nothing to reduce the noise. Due to these reasons, ε value should be set adequately.

4 EXPERIMENT

To clarify the adequateness of the proposed method, we conducted the experiments utilizing a speech signal with a noise signal. In the experiments, we changed ε values in TF ε-filter, and investigate the

Table 1: Prepared speech files.

	Word (Meaning)	Gender
Speech 1	Senkai (Turning)	Male
Speech 2	Senkai (Turning)	Female
Speech 3	Koutai (Backdown)	Male



(a) Speech 1. (b) Speech 2.

(c) Speech 3. (d) Noise.

Figure 2: Waveforms of the signals.

relation between $p(X|W)$, and the mean square error (MSE) between the original speech signal $s(k)$ and the filter output $y(k)$. MSE is defined as follows:

$$MSE = \frac{1}{L}\sum_{k=1}^{L}(s(k)-y(k))^2, \qquad (6)$$

where L is the signal length. As sound sources, we prepared three Japanese speech signals as shown in Table 1 and Figures 2(a)-(c). As shown in Table 1, the first one and the second one are the same word but pronounced by different persons to check the robustness of gender difference. The first one and the third one are different words pronounced by the same person to check the robustness of word difference.

The sound of clapping hands is used as the noise signal. The waveform of the noise signal is shown in Fig.2(d). We mixed each signal and the noise in the computer. All the *SNR*s of the mixed signals are 1.4[dB]. *SNR* is defined as follows:

$$SNR = 10\cdot\log_{10}\left(\frac{\sum_{k=1}^{L}s(k)^2}{\sum_{k=1}^{L}n(k)^2}\right). \qquad (7)$$

The sampling frequency and quantization bit rate are set at 16kHz and 16bits, respectively. We set the window size of ε-filter at 61.

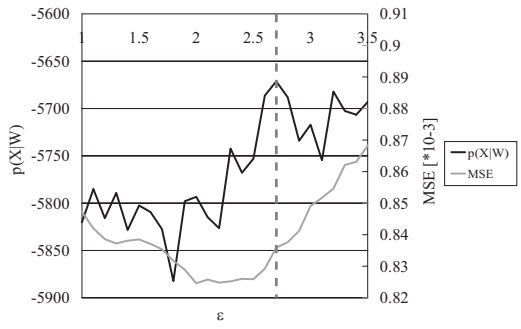


Figure 3: Speech 1.

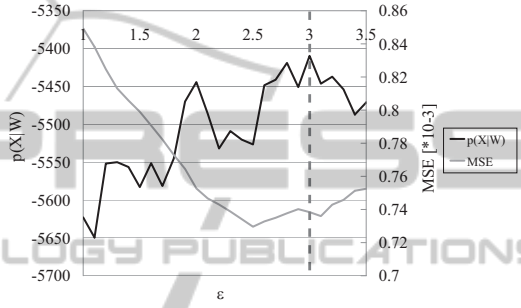


Figure 4: Speech 2.

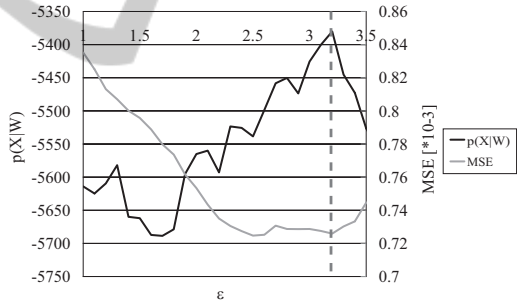


Figure 5: Speech 3.

Figures 3-5 show the experimental results. The dashed line in the figures represents the auxiliary lines, which show the maximal recognition probability.

As shown in Figs.3-5, as is generally considered, MSE became small with regard to ε, which maximizes $p(X|W)$. The proposed criterion is also robust for gender difference and word difference.

We also note that the parameter that minimizes MSE does not exactly correspond to the parameter that maximizes $p(X|W)$. It is considered that speech recognition system is biased by auditory model obtained through learning process. Although many researchers currently aim to eliminate such biases, and evaluate the system performance by objective functions, we think that it is necessary to take these types

of biases into consideration when we apply the filtering system to recognition system.

5 DISCUSSIONS AND CONCLUSIONS

In this paper, we proposed parameter setting of noise reduction filter utilizing speech recognition system. The algorithm is simple, and the adequate parameter could be obtained throughout the experiments. As the proposed method does not use the relation between the signal and noise, it is expected that the application range of the proposed method is large. By using our method, even if we only have the single-channel noisy signal, we can evaluate whether the parameter is adequate or not. The proposed method does not require to estimate the noise in advance.

Experimental results also give us some visions to be considered with regard to our approach. For instance, the nonlinear relation between the filter parameter and the recognition result is an important problem. The obtained recognition result sometimes drastically moves with a tiny parameter change due to the nonlinearity between the filtering and recognition system. Our approach will be more useful when a filter has the linear relationship between the parameter change and the filtering error.

For future works, we would like to investigate the relation between the adequate recognition system and filtering system. Theoretical analyses are also required. We aim to apply this criterion to other systems. Applications for robot auditory will also be considered.

ACKNOWLEDGEMENTS

This research was supported by Special Coordination Funds for Promoting Science and Technology, by research grant from Support Center for Advanced Telecommunications Technology Research (SCAT), by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 22700186, 2010. and by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 20700168, 2008. This research was also supported by the CREST project “Foundation of technology supporting the creation of digital media contents” of JST, and the Global-COE Program,“Global Robot Academia”, Waseda University.

REFERENCES

Abe, T., Matsumoto, M., and Hashimoto, S. (2007). Noise reduction combining time-domain ε-filter and time-frequency ε-filter. In *J. of the Acoust. Soc. America.*, volume 122, pages 2697–2705.

Abe, T., Matsumoto, M., and Hashimoto, S. (2009). Parameter optimization in time-frequency ε-filter based on correlation coefficient. In *Proc. of SIGMAP2009.*

Harashima, H., Odajima, K., Shishikui, Y., and Miyakawa, H. (1982). ε-separating nonlinear digital filter and its applications. In *IEICE trans on Fundamentals.*, volume J65-A, pages 297–303.

Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada, A., Utsuro, T., and Shikano, K. (2000). Free software toolkit for japanese large vocabulary continuous speech recognition. In *Proc. of ICSLP 2000.*

Lee, A. and Kawahara, T. (2009). Recent development of open-source speech recognition engine julius. In *Proc. of APSIPA-ASC 2009.*

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius — an open source real-time large vocabulary recognition engine. In *Proc. of EUSIPCO2001.*, pages 1691–1694.

Lim, J. S., Oppenheim, A. V., and Braida, L. (1978). Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition. In *IEEE Trans. on Acoust. Speech Signal Process.*, volume ASSP-26, pages 419–423.

Matsumoto, M. and Hashimoto, S. (2009). Estimation of optimal parameter in ε-filter based on signal-noise decorrelation. In *IEICE Trans. on Information and Systems*, volume E92-D, pages 1312–1315.

Mrazek and Navara, M. (2003). Selection of optimal stopping time for nonlinear diffusion filtering. In *Int’l Journal of Computer Vision*, volume 52, pages 189–203.

Sporring, J. and Weickert, J. (1999). Information measures in scale-spaces. In *IEEE Transs on Information Theory*, volume 45, pages 1051–1058.

Umeda, S. (2000). Blind deconvolution of blurred images by using ica. In *IEICE Trans. on Fundamentals*, volume J83-A, pages 677–685.