社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-4

Robust Speech Recognition Using Optimized Wavelet Filtering in Reverberant Conditions

Randy Gomez and Tatsuya Kawahara
Kyoto University,
Academic Center for Computing and Media Studies (ACCMS),
Sakyo-ku, Kyoto 606-8501, JAPAN

Abstract

Speech recognition in reverberant environments is a difficult task. Reverberation has the effect of degradation of recognition performance due to acoustic mismatch. We present an optimization method of the wavelet parameters for dereverberation in automatic speech recognition (ASR). By tuning the wavelet parameters to improve the acoustic model likelihood, wavelet-based dereverberation methods become more effective in the ASR application. We evaluate several existing wavelet-based methods and optimize them, based on our proposed scheme. Experimental evaluations through ASR experiments demonstrate significant improvement for all methods with the proposed optimization.

Index Terms: Robustness, Speech recognition, Dereverberation

1 Introduction

Reverberation is a phenomenon caused by the reflection of the speech signal in an enclosed environment. When analyzing in short time fourier transform (STFT), the current observed speech frame is smeared with the speech energy of the preceding frames. This degrades the acoustic quality of the speech signal and is detrimental to the ASR system. The reverberant speech model $X(f,t)$ we adopt is based on the additive effects of the early $X_E(f,t)$ and late $X_L(f,t)$ reflection,

$$\begin{aligned} X(f,t) &\approx X_E(f,t) + X_L(f,t) \\ &\approx S(f,t)H(f,0) + \sum_{d=1}^{D} S(f,t-d)H(f,d) \end{aligned} \quad (1)$$

where $S(f,t)$ and $H(f,t)$ are the frequency response of the clean speech and the room impulse response (RIR), respectively. D is the number of frames, over which the reverberation (smearing) has an effect. The early reflection is due to the direct signal and some reflections that occur at earlier time, while the late reflection, whose effect spans over frames, can be treated as long-period noise [1]-[4]. The former is mostly addressed through Cepstral Mean Normalization (CMN) in the ASR system as it falls within the frame. In our application, dereverberation is defined as suppressing the effects of the late reflection. Since the late reflection can be treated as noise, we can apply existing wavelet-based denoising techniques to dereverberation problems based on the context of our reverberant speech model.

Most of the speech enhancement algorithms are applied in the frequency domain, using short-time Fourier transform (STFT) where the time resolution is the same for all frequency components. Some enhancement methods are applied in wavelet domain which provides more flexible time-frequency representation of speech. There have been a lot of research involving wavelet-based speech enhancement primarily in denoising [5]-[8]. Originally, wavelet-based enhancement methods were proposed to address denoising problems. Most recently, it is expanded to address the effects of reverberation.

Existing wavelet-based methods are generally designed to enhance the speech waveform, but this does not guarantee an improvement in performance for ASR application. In this paper, we present a method of optimizing the wavelet parameters for dereverberation in ASR. In our proposed scheme, prior to wavelet-based dereverberation, the wavelet parameters are optimized to improve the likelihood of the acoustic model. We expand existing wavelet-based speech enhancement methods for the dereverberation application. Then, we incorporate the proposed scheme of optimizing the wavelet parameters for effective dereverberation in the ASR application. In this paper, noise and late reflection are jointly referred to as "contaminant signal".

The paper is organized as follows; Section 2 gives the background of the different wavelet-based methods which we will evaluate and optimize. In Section 3, we present the optimization method of wavelet parameters. Experimental set-up and ASR evaluation results are pre-

Randy Gomez is a research fellow of the Japan Society for the Promotion of Science (JSPS).

sented in Section 4. Finally, we conclude this paper in Section V.

2 Dereverberation Methods using Wavelets

In this section, we will discuss existing wavelet-based methods. Specifically in this paper, we consider five wavelet-based methods. The last method was previously proposed by the authors [9].

2.1 WaveShrink

The basic wavelet enhancement approach [10] is based on the idea that real-world signals do not necessarily require full resolution treatment. In speech application, a limited number of wavelet coefficients in the lower band are deemed sufficient to reconstruct the speech signal. These coefficients are characterized by higher values compared to the contaminant signals (i.e. noise or late reflections). Thus, by shrinking the contaminant wavelet coefficients, its effects are removed. In general, the waveshrink approach is applicable when the contaminant signal is homogeneously concentrated on the other side of the spectrum (e.g. higher frequencies). Problems may arise in ASR applications, because some parts of speech have important information in the higher frequencies (i.e consonants and unvoiced regions).

2.2 Thresholding

An improved version of the waveshrink approach is implemented by means of a thresholding algorithm. Unlike its predecessor, the thresholding approach is more flexible in dealing with the wavelet coefficients by defining a threshold criterion. A particular wavelet coefficient of interest may be shrunk or scaled based on this criterion. An example based on soft thresholding [11] is defined as

$$\bar{x} = \begin{cases} 0 & ,\mid x \mid \leq \ thr \\ sign(x)(\mid x \mid - thr) & ,\mid x \mid > \ thr \end{cases} \tag{2}$$

Based on the threshold *thr*, Eq. (2) can be interpreted as setting the contaminant subspace to zero, and implementing a magnitude subtraction in the speech plus contaminant subspace. The threshold that defines the subspace of the contaminant signal can be calculated [11] as

$$thr = \sigma\sqrt{2 \ log(L)}, \tag{3}$$

where L is the length of the contaminant signal with variance σ^2. Other thresholding criteria are *Hard*, *Firm*, *Garrote* and $Step - garrote$. The thresholding technique has some known problems; If the spectrum of the contaminant signal is not uniform, the method has difficulty in distinguishing the desired subspace from the contaminant subspace. Since thresholding is directly applied to the wavelet coefficients, the quality of the reconstructed signal is sensitive to the threshold.

2.3 Improved Wavelet-based Speech Enhancement System

To address the problems in both the waveshrink and thresholding methods, a more advanced method is proposed [12]. This system employs an automatic pause detection algorithm using a voice activity detection (VAD) and introduces several threshold profiles for different types of contaminant signals. With the VAD, a more accurate estimation of noise power is achieved. In addition to the VAD, it incorporates speech signal features in the system. It also implements a mechanism that efficiently selects suitable parameters for voiced, unvoiced and silence regions, separately. The use of several threshold profiles enables switching several threshold criteria according to the contaminant signal. Consequently, the system can cope with colored and non-stationary contaminant signals.

2.4 Wavelet Extrema Clustering

Another method based on the adoption of the speech production model is the wavelet extrema clustering. It assumes that the detrimental effects of the contaminant signal introduce zeros into the overall system and only affects the speech excitation sequence (not the all-pole filter) [13]. A class of wavelets are employed to decompose the LPC residuals to calculate the wavelet extrema. The underlying impulsive structure of the desired speech (non-reverberant) are captured by locating the extrema which has the characteristics of being well clustered. The extrema at each wavelet scale are effective indicators of the impulses (clean speech) in the contaminated signal. These are used to reconstruct the non-reverberant speech.

2.5 Wavelet Filtering with Wiener Gain

We have previously expanded the multi-band wavelet domain filtering [9] to address the dereverberation problem [14]. The general expression of the Wiener gain at band m [14] is expressed as

$$\kappa_m = \frac{S(\upsilon,\tau)_m^2}{S(\upsilon,\tau)_m^2 + X_L(\upsilon,\tau)_m^2}, \tag{4}$$

where $S(\upsilon,\tau)_m^2$ and $X_L(\upsilon,\tau)_m^2$ are wavelet power estimates for the clean speech and the late reflection, respectively. And υ and τ are the wavelet parameters scale and shift, which will be explained in Section 3. Wavelet filtering is carried out by weighting the reverberant wavelet coefficients $X(\upsilon,\tau)$ with the Wiener gains as,

$$X(\upsilon,\tau)_m(enhanced) = X(\upsilon,\tau)_m \ . \ \kappa_m. \tag{5}$$

In Eq. (5), the Wiener weighting κ_m dictates the degree of suppression of the late reflection to the observed signal. If the late reflection 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 the late reflection. Moreover, 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

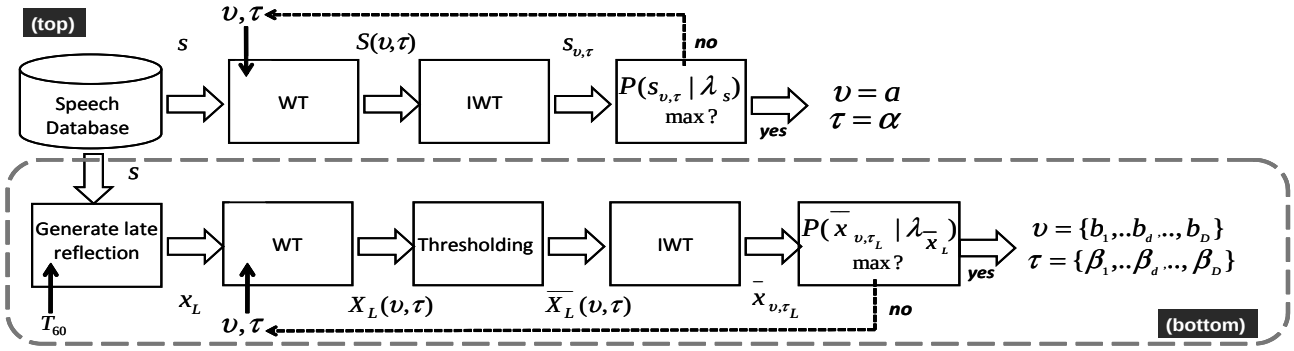

Figure 1: Block diagram of the wavelet optimization scheme.

inverse wavelet transform (IWT). In our previous work [14], the wavelet parameters are not optimized to track the clean speech and the late reflection given a reverberant observation.

3 Optimizing wavelet Parameters υ and τ based on Acoutsic Model Likelihood

A wavelet is generally expressed as

$$\Psi(\upsilon,\tau,t) = \frac{1}{\sqrt{\upsilon}}\Psi\left(\frac{t-\tau}{\upsilon}\right), \tag{6}$$

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

$$F(\upsilon,\tau) = \int f(t)\Psi(\upsilon,\tau,t)dt, \tag{7}$$

where $F(\upsilon,\tau)$ is the wavelet coefficients 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.

For illustration purpose, we will only show the optimization of the wavelet parameters υ and τ for the wavelet filtering method discussed in Section 2.5. In the wavelet filtering method, we are interested in detecting the power of clean speech and late reflection given a reverberant signal.

We optimize the wavelet to detect clean speech and late reflection separately based on the acoustic model likelihood as shown in Fig. 1. In ASR, we assume that the speech does not vary for a certain time-frame. Thus, optimizing a single wavelet template for speech will be sufficient. In Fig. 1 (top) we illustrate the optimization of the wavelet for clean speech. Wavelet coefficients $S(\upsilon,\tau)$, extracted through Eq. (7), are converted back to time domain $s_{\upsilon,\tau}$. Likelihood scores are computed using the clean speech acoustic model λ_s. The process is iterated, adjusting υ and τ. The corresponding υ=a and τ=α that result to the highest score are selected. In the case of the late reflection in Fig. 1 (bottom), D templates are to be optimized for both scale $(\upsilon_1,...\upsilon_D)$ and shift $(\tau_1,...,\tau_D)$. These correspond to D preceding frames that cause smearing to the current frame of interest. We note that the effect of smearing is not constant, thus D templates are created. By estimating the reverberation time T_{60}, we can generate the impulse response and its corresponding late reflection coefficients h_L. Both T_{60} estimation and impulse response generation are discussed in [15]. Then, late reflection observations x_L are generated by convolving the clean speech with h_L. Next, wavelet coefficients $X_L(\upsilon,\tau)$ are extracted through WT (Eq. (7)). To make sure that $X_L(\upsilon,\tau)$ is void of speech characteristics, thresholding is applied to $X_L(\upsilon,\tau)$. Speech energy is characterized with high coefficient values [11] [12] and thresholding sets these coefficients to zero,

$$\bar{X}_L = \begin{cases} 0 & ,\mid X_L \mid > \ thr \\ X_L & ,\mid X_L \mid < \ thr \end{cases} \tag{8}$$

thr is calculated similar to that in Eq. (3). The thresholded signal is converted back to time domain $\bar{x}_{\upsilon,\tau_L}$ and evaluated against a late reflection model $\lambda_{\bar{x}_L}$. The parameters υ and τ are adjusted and the corresponding υ=$\{b_1,...b_D\}$ and τ=$\{\beta_1,...\beta_D\}$ that result to the highest likelihood score are selected. We note that the acoustic model λ_s is trained with clean speech data, while $\lambda_{\bar{x}_L}$ uses the synthetically generated late reflection data with thresholding applied.

By using these optimized wavelet parameters, we can estimate both the clean speech and late reflection power directly from the observed reverberant signal $X(\upsilon,\tau)$ and use these to estimate the Wiener gain in Eq. (4). Thus, the speech power estimate becomes

$$S(\upsilon,\tau)_m^2 \approx X(a,\alpha)_m^2, \tag{9}$$

and the late reflection power $X_L(\upsilon,\tau)_m^2$ estimate

$$X_L(b_d,\beta_d)_m^2 \approx \begin{cases} X(b_1,\beta_1)^2 \ , \ d=\ 1 \\ \dfrac{\sum_{k=1}^{d-1} X(b_k,\beta_k)^2}{d-1} + X(b_d,\beta_d)_m^2 \ , \\ \text{otherwise} \end{cases} \tag{10}$$

where d (smearing effect) is the d-th frame template (for k:1,...,D).

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	8256 Gaussian pdfs
Training data	Adult by JNAS
Test data	Adult by JNAS

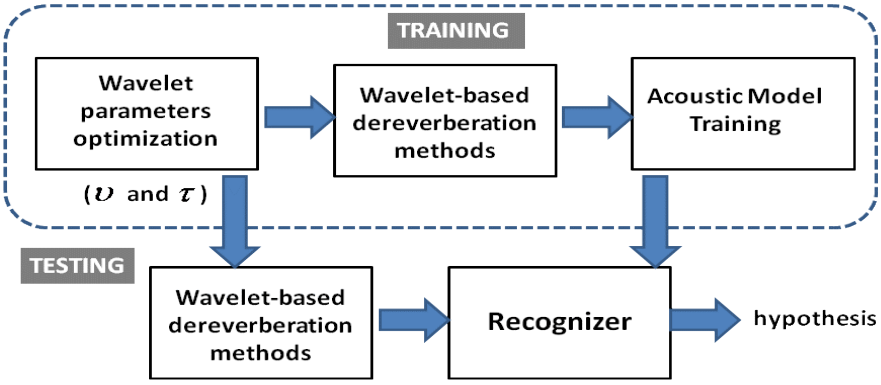

Figure 2: Overall system diagram (Training and Testing).

Table 2: Recognition performance for different wavelet-based methods (No adaptation).

	200 ms	**400 ms**	**600 ms**	**average**
No processing; clean model	68.6 %	41.3 %	21.4 %	43.8 %
No processing; reverberant model	75.4 %	61.2 %	32.1 %	56.2 %
(1) WaveShrink (Sec. 2.1)	75.9 %	63.3 %	40.6 %	60.0 %
(1+) WaveShrink + wavelet optimization	76.7 %	65.4 %	44.9 %	62.3 %
(2) Soft thresholding (Sec. 2.2)	76.5 %	65.8 %	46.7 %	63.0 %
(2+) Soft thresholding + wavelet optimization	78.1 %	67.1 %	49.2 %	64.8 %
(3) Improved wavelet-based speech enhancement (Sec. 2.3)	77.3 %	66.7 %	50.6 %	64.8 %
(3+) Improved wavelet-based speech enhancement + wav. opt.	79.1 % %	68.5 %	54.0 %	67.2 %
(4) Extrema clustering (Sec. 2.4)	78.4 %	67.1 %	59.7 %	68.4 %
(4+) Extrema clustering + wavelet optimization	80.8 %	69.8 %	62.9 %	71.1 %
(5) Wavelet filtering (Sec. 2.5)	81.5 %	71.4 %	64.5 %	72.5 %
(5+) Wavelet filtering + wavelet optimization	83.2 %	74.6 %	68.6 %	75.5 %

4 Experimental Evaluations

We have evaluated the proposed scheme and the five wavelet-based methods described in Section 2. Evaluation is carried out 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 open test set is composed of 200 utterances uttered by 50 speakers. ASR experiments are carried out on the Japanese dictation task with a 20K vocabulary. The language model is a standard word trigram model. The acoustic model is a phonetically tied-mixture (PTM) HMMs with 8256 Gaussians in total. System specification is summarized in Table 1.

We experimented in the condition of reverberation time: T_{60}=200 ms, 400 ms and 600 ms. Reverberant training data are synthetically produced with the automatically generated RIR as discussed in [15]. Test performance is evaluated using real data recorded in a room with known reverberation time: T_{60}=200 ms, 400 ms and 600 ms. In the experiments, we used a total number of bands $M = 5$ which was found to be effective [1][3]. The wavelet used here is the Daubechies wavelet which was also used in [14].

Table 3: Recognition performance for different wavelet-based methods (MLLR adaptation).

	200 ms	**400 ms**	**600 ms**	**average**
No processing; clean model	70.3 %	43.2 %	24.8 %	46.1 %
No processing; reverberant model	76.5 %	63.2 %	35.1 %	58.2 %
(1) WaveShrink (Sec. 2.1)	76.4 %	64.8 %	41.1 %	60.8 %
(1+) WaveShrink + wavelet optimization	77.9 %	67.2 %	46.4 %	63.8 %
(2) Soft thresholding (Sec. 2.2)	77.8 %	67.5 %	47.1 %	64.1 %
(2+) Soft thresholding + wavelet optimization	79.0 %	68.6 %	51.4 %	66.3 %
(3) Improved wavelet-based speech enhancement (Sec. 2.3)	78.5 %	67.9 %	52.1 %	65.1 %
(3+) Improved wavelet-based speech enhancement + wav. opt.	80.0 % %	69.5 %	56.2 %	68.5 %
(4) Extrema clustering (Sec. 2.4)	79.6 %	68.2 %	61.5 %	69.7 %
(4+) Extrema clustering + wavelet optimization	81.5 %	70.7 %	64.1 %	72.1 %
(5) Wavelet filtering (Sec. 2.5)	82.7 %	72.7 %	66.9 %	74.1 %
(5+) Wavelet filtering + wavelet optimization	84.2 %	76.3 %	69.5 %	76.6 %

The process flow of the experiment is shown in Fig. 2. During training, we optimize the wavelet parameters. Using the optimized wavelet parameters, we implemented the wavelet-based dereverberation methods discussed in Section 2, then trained individual acoustic models. During testing, the optimized wavelet parameters were used together with the wavelet-based dereverberation methods to process the reverberant test data. Then, processed data were evaluated in ASR. In our experiments, the actual optimization of the wavelet parameters may vary for each of the different wavelet-based dereverberation methods, depending on individual unique requirements. Nevertheless, the criterion of maximizing the likelihood for the ASR application is maintained for all the methods.

We also implemented a model adaptation based on Maximum Likelihood Linear Regression (MLLR) [16][17]. Model adaptation is used to minimize the mismatch between training and testing conditions. The MLLR adaptation estimates linear transformations for groups of model parameters to maximize the likelihood of the adaptation data. In our adaptation experiment, we used 50 adaptation utterances.

We show the ASR performance in word accuracy for all methods in Tables 2-3. The conventional acoustic model training based on Baum-Welch is used in Table 1 (No adaptation). In Table 2, acoustic model adaptation was implemented using MLLR. In the case of the MLLR, the adaptation data is limited to using only 10 adaptation utterances. In usual case, several adaptation utterances are used (more than 10) for improved performance. In this experiment, we only wanted to verify whether adaptation works in our proposed method.

For reference, we show on the top the results when the reverberant data are not processed and matched against clean and reverberant acoustic models, respectively. We show the results based on waveshrink and thresholding (Sections 2.1 and 2.2) in (1) and (2), respectively. The improvement in (1+) and (2+) from (1) and (2) are the results when the wavelet parameters are optimized. The improved wavelet-based enhancement system that incorporates VAD and threshold profiles (Section 2.3) is shown in (3). In (3+), an improvement in performance is attained when wavelets are optimized as compared to (3). Another method based on extrema clustering (Section 2.4) is provided in (4) together with the optimized wavelet version in (4+). The result of our previous dereverberation approach (Section 2.5) [14] is shown in (5), while the result of incorporating wavelet optimization discussed in Section 3 is given in (5+).

The results in Tables 2-3 show that all the methods (1-5) benefit from the proposed method. By optimizing the wavelet parameters, the dereverberation process is more tuned to improving the acoustic model likelihood. As a result, it becomes more effective in the ASR application. Moreover, we observe a consistent improvement in recognition performance when the model adaptation was conducted. Thus, the proposed optimized dereverberation method also works in the context of adaptation.

We note that in (1),(2) and (3), dereverberation is implemented by means of directly thresholding the wavelet parameters. This may have detrimental effects to the speech recognition performance due to the non-smooth nature of the thresholding function. In our method, thresholding is only used to select the the optimal wavelet parameters and not directly applied to the wavelet coefficients. The actual weighting of the wavelet coefficients is through Wiener filtering, which is a smoother weighting function based on the power ratio of the estimated clean speech and late reflection. Moreover, (1),(2),(3),(4) and (5) are originally based on improving the speech quality (hearing) of the dereverberated signal. However, improving the speech quality may not necessarily translate to improvement in ASR performance. Thus, when we optimized the system for ASR, we have achieved improvement in the recognition performance.

5 Conclusion

Wavelet-based speech enhancement approach has been successfully used in addressing denoising problems. Its application has been extended to reverberant scenarios. Although satisfactory improvement in signal-to-noise ra-

tio has been reported, the existing approach is primarily optimized for improved human perception. In our method, we are interested in optimizing the wavelet-based dereverberation for ASR.

We proposed to optimize the wavelet parameters used in dereverberation in ASR. This scheme guarantees that the optimized parameters improve the model likelihood used in ASR. We have evaluated existing wavelet-based methods. Moreover, we have shown that our approach is effective in improving the ASR performance when applied to different wavelet-based dereverberation methods. In the future, we investigate the effects of contaminated noise and extend this work to deal with both noisy and reverberant environment conditions.

References

[1] 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 Conference on Acoustics Speech and Signal Processing ICASSP*, 2008.

[2] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop on Hands-free Speech Communication and Microphone Arrays HSCMA*, 2008

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

[4] R. Gomez and T. Kawahara, "Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood" *In Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[5] Vaseghi SV., "Advanced Digital Signal Processing and Noise reduction" *2nd ed. Wileys*, 2005.

[6] Q. Fu and EA. Wan, "Perceptual Wavelet Adaptive Denoising of Speech" *In Proceedings of EUROSPEECH*, 2003.

[7] JW. Seok and KS. Bae, "Speech Enhancement with Reduction of Noise in the Wavelet Domain" *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, 1997.

[8] M. Jansen, "Noise Reduction by Wavelet Thresholding" *In Proceedings of the IEEE International Symposium on Industrial Electronics, ISIE*, 2001.

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

[10] H.Y. Gao, "wavelet Shrinkage Denoising", *Computational Graphical Statistics* 1998.

[11] D.L. Donoho, "Denoising by soft thresholding", *IEEE Trans. Info. Theory* 1995.

[12] H. Sheikhzadeh and Hamid. Abutalebi, "An Improved Wavelet-based Speech Enhancement System" *In Proceedings Eurospeech*, 2001.

[13] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation" *In Proceedings of the IEEE Workshop on Acoustic Echo and Noise Control*, 1999

[14] R. Gomez and T. Kawahara, "Optimizing Spectral Subtraction and Wiener Filtering for Robust Speech Recognition in Reverberant and Noisy Conditions" *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, 2010.

[15] R. Gomez, T. Kawahara, "Tight Integration of Dereverbeartion and Automatic Speech Recognition" *In proceedings of the Asia Pacific Signal and Information Processing Association APSIPA*, 2009.

[16] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework" *In Proceedings of Computer Speech and Language*, 1996.

[17] Leggeter, C.J., Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models" *In Proceedings of Computer Speech and Language*, 1995.