Dereverberation based on Wavelet Packet Filtering for Robust Automatic Speech Recognition

Randy Gomez and Tatsuya Kawahara

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

Abstract

This paper describes a multiple-resolution signal analysis to suppress late reflection of reverberation for robust automatic speech recognition (ASR). Wavelet packet tree (WPT) decomposition offers a finer resolution to discriminate the late reflection subspace from the speech subspace. By selecting appropriate wavelet basis in the WPT for speech and late reflection, we can effectively estimate the Wiener gain directly from the observed reverberant data. Moreover, the selection procedure is performed in accordance with the likelihood of acoustic model used by the speech recognizer. Dereverberation is realized by filtering the wavelet packet coefficients with the Wiener gain to suppress the effects of the late reflection. Experimental evaluations with large vocabulary continuous speech recognition (LVCSR) in real reverberant conditions show that the proposed method outperforms conventional wavelet-based methods and other dereverberation techniques.

Index Terms: Speech recognition, Robustness, Dereverberation, Wavelet Packets

1. Introduction

In reverberant environments, smearing of the observed signal by the effects of reflection causes acoustic model mismatch. Dereverberation methods based on the suppression of late reflection have been proposed [1][2]. An expansion to this work using multi-band processing is also proposed [3]. In these methods [1]-[3], it was established that the effects of late reflection is more detrimental to ASR. In general, estimating late reflection is pivotal to the effectiveness of discriminating its subspace. However, the estimation is difficult especially if the late reflection subspace overlaps with speech. We have previously proposed a wavelet filtering approach based on pre-determined bands [4] . Although this method works well, fixing the bands limit the ability for the wavelet parameters to effectively capture the subspaces for both speech and late reflection especially during mismatch conditions.

In this paper, we address the subspace discrimination problem with more precision through wavelet packet (WP) analysis. Late reflection is suppressed by filtering the reverberant wavelet packet coefficients with a Wiener gain. Wavelet packet tree (WPT) decomposition is optimized using acoustic model likelihood criterion to effectively track both speech and late reflection, resulting to an accurate Wiener estimate. The WPT which contains the wavelet basis is kept for the actual online dereverberation. Fig. 1 shows the online dereverberation scheme. First, the room reverberation time T_{60} is estimated. The corresponding WPTs are used to decompose the reverberant speech resulting to three WPT decompositions (i.e. late reflection, speech, reverberant speech). Through the WP analysis using the wavelet basis associated with the WPT, WP coefficients are calculated

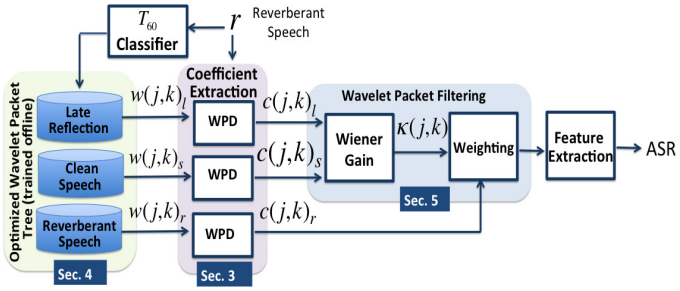

Figure 1: Block diagram of the proposed method.

and used in WP filtering. Finally, the enhanced signal is further processed to extract features for ASR. Although WP analysis has been primarily studied for speech enhancement, our study is focused on its tight integration with ASR.

The paper is organized as follows; Section 2 shows the concept of the dereverberation approach we adopt. In Section 3, wavelet analysis through WPT is introduced. In Section 4, we present the method of selecting appropriate WPT based on the entropy and acoustic model likelihood criterion. The actual dereverberation based on WP filtering is described in Section 5, followed by the experimental results in Section 6. Finally, we conclude this paper in Section 7.

2. Dereverberation Concept

We denote the spectral feature (f:frequency, m:frame) of the reverberant signal, clean speech signal, and room impulse response (RIR) as $R(f,m)$, $S(f,m)$ and $H(f,m)$, respectively. The reverberant speech model [3] expressed in terms of early and late reflections is approximated as

$$\begin{array}{ccccc} R(f,m) & \approx & S(f,m)H(f,0) & + & \sum_{d=1}^{D} S(f,m-d)H(f,d) \\ & \approx & E(f,m) & + & L(f,m) \end{array} \quad (1)$$

where $H(f,0)$ is the RIR effect to the speech signal $S(f,m)$ attributing to the early reflection $E(f,m)$. The second term $L(f,m)$ referred to as late reflection can be viewed as smearing of the clean speech by $H(f,d)$ which corresponds to the d frame-shift effect of the RIR. D is the number of frames over which the reverberation has an effect. The early reflection is mostly addressed through Cepstral Mean Normalization (CMN) in ASR. Therefore, dereverberation is reduced to suppressing the effects of the late reflection $L(f,m)$. Since the late reflection can be treated as additive noise formulated in Eq. (1), dereverberation is simplified to a denoising problem.

3. Wavelet Packet Tree (WPT) Analysis

A one-dimensional wavelet is generally expressed as

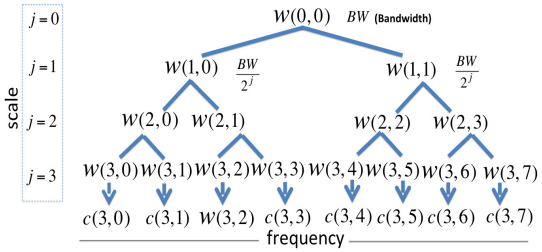

Figure 2: Wavelet Packet Tree (WPT) decomposition.

$$\Psi_{j,k}(t) = 2^{\frac{-j}{2}} \Psi(2^{-j}t - k), \quad j \in Z \quad k \in Z \quad , \tag{2}$$

where t denotes time, j is the depth of the dyadic scale having a resolution of 2^{-j}, and k is the dyadic translation. Wavelet analysis offers a flexibility of scaling and translating the wavelets which controls the degree of representing signals of interest. A proper choice of these parameters would lead to a better representation of signals.

Fig. 2 shows the WPT decomposition method. The scale j and translation k correspond to the depth and position of the wavelet packets $w(j,k)$ in the tree structure. For a WPT decomposition W, there exists a library of wavelet packets $w(j,k)$, and for every wavelet packet, a wavelet basis $\Psi_{j,k}$ is associated to it. The wavelet basis contains the orthogonal filter information (i.e. high and low pass filters). Every wavelet packet splits the bandwidth of the signal, and as this process continues down the tree, the frequency resolution is further refined. Thus, WPT decomposition is analogous to filterbank analysis. The output of the WP analysis denoted as $c(j,k)$ are called WP coefficients, and from this, together with the wavelet basis, we can synthesize the original signal x,

$$x = \sum_{j \in Z} \sum_{k \in Z} c(j,k) \Psi_{j,k}(t). \tag{3}$$

The use of the tree structure decomposition allows us to analyze the signal of interest in finer resolution by splitting further the tree nodes. Although the WP method trades time resolution with frequency resolution, time is already set when selecting the window frame for the ASR. Thus, the frequency resolution is significant in our application.

4. Selecting Wavelet Basis Function

With an appropriate training algorithm, we can select j and k in the WPT decomposition to capture specific characteristics of a certain signal of interest. The resulting wavelet packets are sensitive in detecting the presence of this signal given any arbitrary signal. In our case, we are interested in detecting the power of speech and late reflection given an observed reverberant signal to effectively estimate the Wiener gain.

4.1. Entropy-based Decomposition

There exists at least $2^{N/2}$ binary subtrees in a complete binary tree decomposition of a signal with N samples, which may be a very large number. To control the splitting of the nodes, we use an entropy-based criterion. We note that over-splitting may result to a large number of nodes containing insignificant information. The resulting leaves of the tree structure represent the spectral distribution of the signal of interest. For typical speech signal, WPT should have more frequency resolution in the lower frequency spectrum in which the speech energy is concentrated, as depicted in Fig. 3. There exist several entropy-based criteria as follows [6].

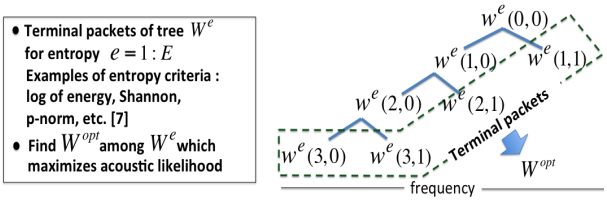

Figure 3: Example of WPT decomposition.

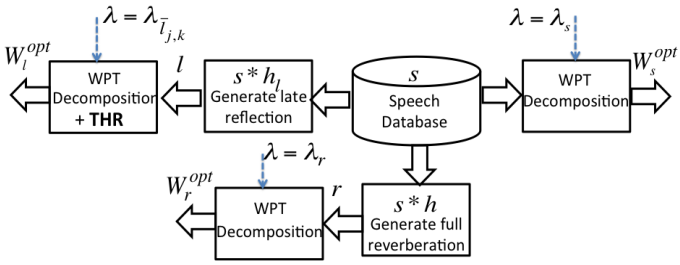

Figure 4: Training wavelet parameters.

- logarithm of the energy entropy:

$$E_{j,k} = \sum_i \log(x_i^2). \tag{4}$$

- Shannon entropy:

$$E_{j,k} = -\sum_i x_i^2 \log(x_i^2). \tag{5}$$

- p norm entropy:

$$E_{j,k} = \sum_i |x_i|^p \qquad p = 1, 2 \text{ and } 3. \tag{6}$$

Here, $E_{j,k}$ is the entropy at each packet and x_i are the coefficients of the signal x in an orthonormal basis at node (j,k). In each splitting process, the cumulative entropy of two split packets are compared with the entropy of its source node. Splitting terminates when the cumulative entropy falls below the entropy of the source node. Specifically, the entropy-based decomposition shown in Fig. 3 is realized as follows.

- *At node (j,k), calculate the cumulative entropy of the split packets* $w(j+1,u)$ *and* $w(j+1,u+1)$*:*

$$Cum_{j+1,k} = E_{j+1,u} + E_{j+1,u+1} \tag{7}$$

- *if* $Cum_{j+1,k} > E_{j,k}$ *then split the node*
- *else terminate, resulting to terminal packet* $\{w(j,k)\}$

The splitting can be conducted using several entropy criteria, including the three listed above ($e = 1 : E$) resulting to $W^{e=1:E}$ tree structures. Using the terminal packets of these tree structures, we search for the best tree W^{opt} among $W^{e=1:E}$ by

- Synthesizing the signal x^e for each tree in $W^{e=1:E}$ using Eq. (3) (Note that this is possible since each packet contains the wavelet basis $\psi_{j,k}$ and the coefficient $c(j,k)$.)
- Select optimal choice W^{opt} by evaluating

$$opt = \arg\max_e P(\mathbf{x}^e|\lambda), \tag{8}$$

where λ is the acoustic model.

We search ($e = 1 : E$) entropy-based criteria for speech, late reflection and the reverberant signal, respectively.

4.2. Training WPT for Speech and Late Reflection

For speech, a single WPT to capture the general speech characteristics is sufficient since we are interested in the speech subspace in general. In Fig. 4, we illustrate the method of selecting the wavelet packets for clean speech. From the clean speech database, WPT is trained as described in Section 4.1. For the acoustic model λ_s, a Gaussian Mixture Model (GMM) of 64 components is used. This is a text-independent model which captures the statistical information of the speech subspace. Specifically, when using speech data s Eq. (8) becomes

$$opt = \arg\max_{e} P(\mathbf{s}^e|\lambda_s),$$

and the resulting tree structure W_s^{opt} is kept.

For late reflection, we discretize T_{60} from 100 ms to 600 ms with 50 ms interval. WPT selection is conducted for each of these. By using the method of T_{60} estimation and synthetic impulse response generation [5][8][3], we can identify the reverberation time T_{60} among the discretized values mentioned above. Consequently, we can generate the RIR h (time-domain equivalent of H in Eq. (1) and its corresponding late reflection coefficients h_l [3]. Then, late reflection observations l are synthetically generated by convolving the clean speech with h_l. Next, WPT is trained in the same manner as in the clean speech, except that thresholding is applied to the WP coefficients prior to synthesis. This ensures that the coefficients are void of speech characteristics. Speech energy is characterized with high coefficient values [7] and thresholding sets these coefficients to zero.

$$\bar{c}(j,k)_l = \begin{cases} 0 & ,\mid c(j,k)_l \mid > \ thr \\ c(j,k)_l & ,\mid c(j,k)_l \mid \leq \ thr \end{cases} \quad (9)$$

The thresholded coefficient is synthesized (Eq. 3) back to time domain $\bar{l}^e_{j,k}$ and evaluated against a late reflection model $\lambda_{\bar{l}}$. Specifically, Eq. (8) becomes

$$opt = \arg\max_{e} P(\bar{\mathbf{l}}^e|\lambda_{\bar{l}}),$$

and the corresponding WPT W_l^{opt} that result to the highest likelihood score is kept. $\lambda_{\bar{l}}$ is trained using the automatically generated late reflection data with thresholding applied.

The proposed WPT selection makes the signal subspaces of speech and late reflection to be effectively discriminated from each other. Thus, W_s^{opt} and W_l^{opt} are of different tree structures. We note that this is not true when simply using very high-resolution filterbanks in which subspaces of speech and late reflection are overlapped, resulting to poor power estimates.

5. Wavelet Packet Filtering

WP filtering is conducted framewise by weighting the contaminated WP coefficient $c(j,k)_r$

$$c(j,k)_{enhanced} = c(j,k)_r \ . \ \kappa(j,k), \quad (10)$$

where the Wiener gain $\kappa(j,k)$ dictates the degree of suppression of the late reflection to the observed signal. The general expression of the Wiener gain is given as

$$\kappa(j,k) = \frac{c(j,k)_s^2}{c(j,k)_s^2 + c(j,k)_l^2}, \quad (11)$$

where $c(j,k)_s^2$, and $c(j,k)_l^2$ are the power estimates for the clean speech and late reflection, respectively. Specifically, these are the WP coefficients of the clean speech s and late reflection l. However, we do not have access to both s and l in the real scenario, but only to the observed reverberant signal r. By using the appropriate WPT decomposition W_s^{opt} and W_l^{opt}, we can estimate the speech power

$$c(j,k)_s^2 \approx c(j_s^{opt}, k_s^{opt})_r^2, \quad (12)$$

where j_s^{opt} and k_s^{opt} are the tree depth and position in the W_s^{opt} decomposition structure. The power estimate of the late reflection is given as

$$c(j,k)_l^2 \approx \frac{1}{D}\sum_{d=1}^{D} \epsilon_d \ . \ c_d(j_l^{opt}, k_l^{opt})_r^2, \quad (13)$$

where $c_d(j_l^{opt}, k_l^{opt})_r^2$ are the estimates for the previous d frames ($d = 1, ..., D$) (see Section 2). ϵ_d is the exponential decay of the reflection energy in the previous d frames [8] which was experimentally derived in [3]. The summation over d represents the smearing effect of the previous frames to the current frame. The left side of Eqs. (12)-(13) are the speech and late reflection power using the actual signal s and l, which is unavailable. The right side is the corresponding approximation using the observed reverberant signal r, when decomposed using W_s^{opt} and W_l^{opt}.

In the actual filtering, we use the tree structure of the reverberant signal as shown in Eq. (10). But the tree structures for both speech and late reflection used in calculating the Wiener gain may be of different depth, resulting to a (usually) shorter terminal leaves. We extend the leaves of these two to correspond to the generic structure of the reverberant signal by padding with zeros. Then, Wiener gain filtering is implemented as described above.

6. Experimental Evaluations

We have evaluated the proposed method in a large vocabulary continuous speech recognition (LVCSR) task. 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 25ms-frame with 10ms shift using Daubechies wavelets. From the enhanced signal via WPT decomposition, we reconstruct the time-domain signal and extract features for ASR. The features used are 12-order MFCCs, ΔMFCCs, and ΔPower. The acoustic model is phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total.

Reverberant training data are synthetically produced with the automatically generated RIR as described in [3]. The test data were recorded in a room with known reverberation time: T_{60}=200ms, 400ms and 600ms. Thus, we used actual reverberant data for evaluation. For reference, the recognition performance for clean speech in word accuracy is 94.0%.

6.1. Comparison with Other Methods

The methods compared in Table 1 are as follows;

- (A) Reverberant data (unprocessed) matched against the clean acoustic model.
- (B) Reverberant data (unprocessed) matched against the reverberant acoustic model.
- (C) A single-band dereverberation method combining linear prediction (LP) residual processing and the spectral processing techniques [2].
- (D) Dereverberation based on the multi-band spectral subtraction [3].

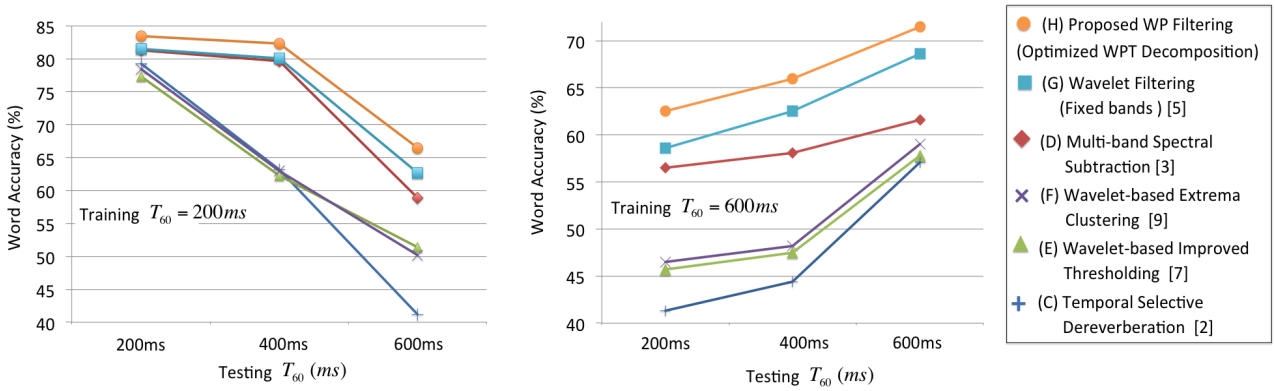

Figure 5: Evaluation against mismatch in reverberant conditions.

Table 1: ASR Result in Word Accuracy (20K LVCSR)

	Real reverberant data		
Methods	200 ms	400 ms	600 ms
(A) No processing; clean model	68.6%	43.1%	21.4%
(B) No processing; reverb. model	72.2%	49.4%	30.3%
(C) Temporal Dereverberation	76.2%	66.0%	57.1%
(D) Multi-band SS	80.7%	71.4%	61.6%
(E) Wavelet-based Thresh.	76.5%	66.2%	57.8%
(F) Wavelet-based Clustering	77.6%	67.9%	59.0%
(G) Wavelet Filtering (fixed)	83.2%	74.5%	68.6%
(H) WP Filtering (Full decomp.)	**83.2%**	**73.3%**	**63.7%**
(I) WP Filtering (Proposed)	**84.5%**	**76.8%**	**71.5%**

- (E) Thresholding in the wavelet domain that incorporates voice activity detection and statistical information [7].
- (F) Wavelet-based method that clusters extrema of the LP coefficients in separating clean components from reverberant components [9].
- (G) Wavelet filtering with pre-defined fixed bands [4].
- (H) WP filtering with conventional WPT full decomposition
- (I) WP filtering with proposed WPT decomposition separately conducted for clean speech, late reflection and reverberant speech.

Table 1 shows the word accuracy for different T_{60}. The acoustic model for each of the methods compared in Table 1 is matched corresponding to the processing of each method. In this table, the proposed method (I) consistently and significantly outperforms other existing methods. Moreover, it is apparent that by using appropriate WPT decomposition (I), an improvement in the recognition performance is achieved from (H). This shows that using different tree structures appropriate for speech and late reflection is more effective than using the simple full WPT decomposition.

6.2. Evaluation in Mismatched Conditions

We investigate the performance of the proposed method in mismatched reverberant conditions. We simulate the mismatched scenario in which the system fails to classify T_{60}. Two models optimized for T_{60} of 200 ms and 600 ms are tested against the data of T_{60} of 200 ms, 400 ms and 600 ms. Fig. 5 demonstrates that the proposed method outperforms the existing methods even in mismatched reverberant conditions. As expected, our previous method of wavelet filtering [4] lags behind the proposed one as the fixed bands cannot cope with the change in reverberant condition.

7. Conclusion

We have proposed a multiple-frequency resolution analysis through the wavelet packets in discriminating the subspaces of clean speech and late reflection. Acoustic likelihood is incorporated with entropy criterion in wavelet basis selection, resulting to a link between the enhancement process and the acoustic model for ASR.

The resultant WPT represents an appropriate frequency resolution of a signal of interest. Therefore, the system can effectively estimate the power of speech and late reflection in reverberant signals. This results to an effective Wiener gain estimate for dereverberation. In the experimental evaluations, the proposed dereverberation method improves ASR performance.

8. References

[1] K. Kinoshita et al., "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *ICASSP*, 2006.

[2] E. Habets, et al., "Temporal Selective Dereverberation of Noisy Speech Using One Microphone" *ICASSP*, 2008.

[3] R. Gomez and T. Kawahara, "Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood" *IEEE Trans. on Audio, Speech and Lang. Proc.*, Sept. 2010

[4] R. Gomez and T. Kawahara, "An Improved Wavelet-based Dereverberation for Robust Speech Automatic Speech Recognition" *Interspeech*, 2010

[5] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise" *SpeCom*, 2008.

[6] R. Gemello et al. "Multiple Resolution Analysis for Robust Automatic Speech Recognition" *Comp. Sp. and Lang.*, 2004.

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

[8] H. Kuttruff, "Room Acoustics" *Spon Press*, 2000

[9] S. Griebel and M. Brandstein, "Wavelet Transform Etxrema Clustering for Multi-channel Speech Dereverberation" *In Proc. IEEE Workshop on Acous. Echo and Noise Control*, 1999.

[10] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" *J. Acous. Soc. of America*, 1995