ADAPTIVE WAVELET PACKET THRESHOLDING WITH ITERATIVE KALMAN FILTER FOR SPEECH ENHANCEMENT

Mengjiao Zhao, Wei-Ping Zhu

Department of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8
Emails: z_mengji@encs.concordia.ca, weiping@ece.concordia.ca

ABSTRACT

In this paper, we propose an adaptive wavelet packet (WP) thresholding method with iterative Kalman filter (IKF) for speech enhancement. The WP transform is first applied to the noise corrupted speech on a frame-by-frame basis, which decomposes each frame into a number of subbands. For each subband, a voice activity detector (VAD) is designed to detect the voiced/unvoiced parts of the speech. Based on the VAD result, an adaptive thresholding scheme is then utilized to each subband speech to obtain the pre-enhanced speech. To achieve a further level of enhancement, an IKF is next applied to the pre-enhanced speech. The proposed method is evaluated under various noise conditions. Experimental results are provided to demonstrate the effectiveness of the proposed method as compared to some previous works in terms of segmental SNR and perceptual evaluation of speech quality (PESQ) as two well-known performance indexes.

Index Terms — Wavelet packet transform, voice activity detection, adaptive wavelet threshold, iterative Kalman filter, speech enhancement.

1. INTRODUCTION

Speech enhancement has been extensively studied for many years and various speech enhancement methods have been developed during the past decades [1]. One of the objectives of speech enhancement is to provide high-quality speech communication in the presence of background noise and concurrent interference signals [2]. In the process of speech communication, the clean speech signal is inevitably corrupted by acoustic noise from the surrounding environment, communication equipments, electrical noise, other speakers, and other sources of interference. These disturbances can significantly degrade the quality and intelligibility of the received speech signal. Therefore, it is of great interest to develop an efficient speech enhancement technique to recover the original speech from noisy observations.

Wavelet Packet (WP), as a well-known powerful method, is used in several signal processing applications including speech enhancement. Based on an adaptive thresholding in wavelet packets, a speech enhancement approach is proposed in [3], where a subband thresholding is applied to detect the voice/noise frames. The criterion used in [3] to determine voiced and noise frames is to compare the frame energy with a constant threshold. Namely, a voiced frame is detected if the frame energy exceeds the constant threshold. Otherwise, the frame is decided as noise. In practice, however, such an energy-based decision making would fail to identify all voice or noise frames. To increase the detection accuracy, other frame characteristics need to be considered.

The authors in [4] proposed an iterative Kalman filter (IKF) approach for speech enhancement, where the linear prediction coefficients (LPCs) and the noise variance are estimated directly from the noisy speech, which decreases the accuracy of IKF approach to a certain degree. In [5], another IKF-based approach is proposed along with a subband processing, where the noisy speech is decomposed into a number of subbands followed by Kalman Filtering (KF) on each subband. The method, however, demands a large amount of computational resource for the implementation of KF at all the subbands. More recently, a subband IKF method with partial reconstruction is proposed in [6], where the noisy speech is first decomposed into a set of subbands, and then a partial reconstruction scheme is used to reconstruct the subbands into high-frequency and low-frequency subband speeches. In this context, the IKF is employed only in the high-frequency subband. Since the low-frequency subband is not filtered by the IKF, this method offers limited enhancement performance for noisy speeches which contain non-negligible noises in low-frequency region.

In order to address the aforementioned limitations, in this paper we propose an improved thresholding scheme with the IKF for speech enhancement on a frame-by-frame basis. The noisy speech is first decomposed into a number of subbands with the WP. The VAD is then applied to each subband frame to determine whether the frame is voice or noise. In contrast to most existing works where only a single parameter is employed for voice/noise frame detection, our method makes use of two measurements in the VAD stage. i) frame energy and ii) spectral flatness. A VAD based adaptive thresholding scheme is then proposed for speech enhancement in accordance with each subband frame activity. Finally, an IKF is used for further noise reduction, which is followed by reconstruction of the full-band speech from the enhanced subband speeches.

The rest of the paper is organized as follows: Section 2 presents the proposed method in detail. In Section 3, we simulate the approach in different noise environments with comparison to other competitive methods. Finally, Section 4 concludes the paper.
2. PROPOSED APPROACH

Consider a time-domain noisy speech \( y(k) \) as given by
\[
y(k) = s(k) + v(k),
\]
where \( s(k) \) is the \( k^{th} \) sample of the clean speech, and \( v(k) \) is the noise sample. In this paper, the input noisy speech is first segmented into frames \( y_n(k) \), where \( n \) is the frame index. The subsequent processing is then carried out on a frame by frame basis. Our proposed approach consists of two successive stages. In the first stage, an improved VAD based adaptive WP thresholding scheme is developed to reduce the noise for the unvoiced frames for each subband. In the second stage, the reconstructed and pre-enhanced full-band speech is processed by the IKF for further enhancement. The details of the proposed method are presented in the following two subsections.

2.1. Subband VAD based adaptive thresholding

In this subsection, \( y_n(k) \) is processed by using the subband VAD scheme along with adaptive WP thresholding. The block-diagram of the pre-enhancement stage is shown in Fig. 1, where the framed speech is first decomposed into several subbands. Each subband speech is described as \( \tilde{y}_i(n)(k) \), where \( i \) is the subband index. A VAD based adaptive thresholding scheme is then applied to each subband, yielding a modified subband speech \( \hat{y}_i(n)(k) \). After processing all the subbands, the WP reconstruction is adopted in order to reconstruct the full-band enhanced speech signal \( \hat{y}_n(k) \).

It is worth mentioning that the number of decomposition channels (subbands) in WP analysis is usually a power of two, which can easily be implemented by several levels of decomposition, each level creating twice subbands. In this paper, we adopt a 3-level WP decomposition, yielding a total of 8 subbands. As each subband \( \hat{y}_i(n)(k) \) goes through the same VAD based thresholding scheme, we will drop the subband index \( i \) in the followig discussions.

Fig. 2 shows the flowchart of the VAD based adaptive thresholding approach. The main idea of the VAD scheme is to extract the measured features from the input noisy speech, and compare them with the feature thresholds, which are computed from noise-only periods. A voiced frame is detected if the measured values exceed the respective threshold. Otherwise, the input speech frame is considered as a noise frame. When the VAD is performed, a voice frame is flagged as VAD = 1, while a noise frame is marked as VAD = 0.

In our VAD scheme, we adopt the energy as the first feature for each frame \( E_n \). The frame energy is calculated by
\[
E_n = \sum_{k=(n-1)L}^{nL} |\hat{y}_n(k)|^2,
\]
where \( L \) is the number of samples in each frame. As the metric (2) ignores the frequency features, it does not provide good feature description for input speech with lower SNRs. Therefore, we would like to take frequency features into account. This feature is called the spectral flatness (\( F \)) [7], which is defined as
\[
F_n = 10 \log_{10}(\frac{m_a}{m_g}),
\]
where \( m_a \) and \( m_g \), respectively, denote the arithmetic and geometric means of the noisy speech spectrum. The spectral flatness is a measure of the noisiness of spectrum and is a good feature in voiced / uncoiced detection [8]. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands, which behaves more like voice frames. However, a high spectral flatness shows that the spectrum power is more uniform in different frequency bands, and appears relatively flat and smooth, which appears more likely as noise.

For the proposed VAD algorithm, we first consider the threshold initialization. For each decomposed subband, we calculate the two features according to (2) and (3) for the first \( N \) frames, then the minimum value of each feature among these frames is taken as the initial thresholding value for the feature as denoted by \( E_{T,0} \) and \( F_{T,0} \) respectively.

The VAD process starts with calculating the two features for frame \( n (n \geq 1) \) obtained from (2) and (3), which results in \( E_n \) and \( F_n \). Both feature values will start to compare with the initial thresholding values \( E_{T,0} \) and \( F_{T,0} \) respectively. As suggested in [8], if the two feature values exceed the thresholds \( E_{T,0} \) and \( F_{T,0} \) respectively, the speech frame \( n \) is marked as a voice frame and the two thresholding values are not updated. Otherwise, frame \( n \) is marked as a noise frame, and the two thresholding values are then updated as
\[
E_{T,n_i} = 40 \log_{10}(\frac{(n_i - 1)E_{T,n_i-1} + E_{n_i}}{n_i}) + E_{T,0},
\]
\[
F_{T,n_i} = \alpha F_{T,n_i-1} + (1 - \alpha)F_{n_i} + F_{T,0},
\]
where \( n_i \) is the index of the noise only frame detected and \( \alpha \) is the exponential smoothing factor. Fig. 3 shows an example of the proposed VAD results. The detected noisy speech has 10 frames and each frame length is \( L = 64 \). As we can see, the frames 1, 2, 6 and 7 are marked as noise frames, while frames 3 − 5 and 8 − 10 are detected as voice frames.

Following the VAD step, the noise and voice frames are
detected. Based on each subband frame activity, the estimated noise variance ($\sigma^2_{v,n}$) and the frame-dependent threshold ($T$) are updated as

$$\sigma^2_{v,n} = \lambda\sigma^2_{v,n-1} + (1-\lambda) \sum_{k=(n-1)L}^{nL} \|\tilde{y}_n(k)\|^2,$$

$$T_n = \lambda T_{n-1} + (1-\lambda) \frac{\text{MAD}_{n1}/0.6745}{\sqrt{2\log_2(L\log_2 L)}},$$

where $\sum_{k=(n-1)L}^{nL} \|\tilde{y}_n(k)\|^2$ denotes the power of the newly detected noise frame. The term $(\text{MAD}_{n1}/0.6745)\sqrt{2\log_2(L\log_2 L)}$ is achieved based on [9], where MAD_{n1} is the the median absolute value of the $n_1$-th detected noise frame. Parameter $\lambda$ is a scaling factor which affects the estimation accuracy. It is noted that $n_1$ is the index of the detected noise frame, which means the values of estimated noise variance ($\sigma^2_{v,n_1}$) and the frame-dependent threshold ($T$) will be updated when the coming frame is a noise frame based on the VAD result, which makes the thresholding value adaptive for each frame. Segmental SNR for each frame is also updated along with the corresponding noise variance $\sigma^2_{v,n}$, which is presented as

$$\text{SNR}_{\text{seg,n}} = 10\log_{10} \sum_{k=(n-1)L}^{nL} \|\tilde{y}_n(k)/\sigma_{v,n}\|^2.$$  

It is noted that segmental SNR reflects noise portion in speech frames, which is an important parameter in deciding the threshold value for each frame. Then an adaptive threshold $\hat{T}_n$ based on each frame segmental SNR is determined as

$$\hat{T}_n = \begin{cases} T_n + T_n e^{-\text{SNR}_{\text{seg,n}}/\tau} & \text{if } \text{SNR}_{\text{seg,n}} \geq 0 \\ 2T_n & \text{if } \text{SNR}_{\text{seg,n}} < 0 \end{cases}$$

where $T_n$ is the threshold value for each frame. The nonlinear function $e^{-\text{SNR}_{\text{seg,n}}/\tau}$ is used to gradually suppress the value of $\hat{T}_n$ when $\text{SNR}_{\text{seg,n}}$ increases where $\tau$ is a constant to set the curve inclination properly. After getting the adaptive threshold value $\hat{T}_n$, the subband speech $\tilde{y}_n(k)$ is compared sample by sample with $\hat{T}_n$ to either suppress the speech values or remain the same values, giving the modified subband speech.

$$\tilde{y}_n(k) = \begin{cases} y_n(k), & |\tilde{y}_n(k)| \geq \hat{T}_n \\ \text{sgn}(k)|\tilde{y}_n(k)|^{3/\hat{T}_n^2}, & |\tilde{y}_n(k)| < \hat{T}_n \end{cases}$$

where $\text{sgn}(k)|\tilde{y}_n(k)|^{3/\hat{T}_n^2}$ denotes a non-linear sign function which is employed to avoid the musical noise. It decreases the WP coefficients values $\tilde{y}_n(k)$ when the sample value is smaller than $\hat{T}_n$. The inverse WP transform is then applied to each subband $\tilde{y}_n(k)$ in order to reconstruct the pre-enhanced full-band speech signal $\hat{y}(k)$.

**2.2. Iterative Kalman Filter**

Here, the pre-enhanced full-band speech signal $\hat{y}(k)$ is further processed by an IKF as modelled below

$$\hat{y}(k) = Hx(k) + w(k),$$

$$x(k) = Fx(k-1) + Gu(k),$$

with $H = G^T = [1, \ldots, 1] \in \mathbb{R}^{1 \times p}$, $x(k) = [s(k-p+1), \ldots, s(k)]$. The term $F$ denotes the $p \times p$ state transition matrix represented as LPCs estimation based on Modified Yule-Walker equations [10]

$$F = \begin{bmatrix} a_1 & -a_2 & \cdots & -a_{p-1} & -a_p \\ 1 & 1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$  

The number of iterations is usually set to 2 or 3. The operation principle of the IKF includes a prediction step and a measurement update step. In the prediction step, the IKF predicts the state vector and parameter covariance by using the previous samples of the state-space model. The estimate of clean speech $\hat{x}(k)$ and the posteriori estimation error covariance $P(k|k)$ are predicted from time step $(k-1)$ to step $k$ (the status are $\hat{x}(k|k-1)/P(k|k-1)$)

$$\hat{x}(k|k-1) = F\hat{x}(k-1|k-1),$$

$$P(k|k-1) = FP(k-1|k-1)F^T + G\sigma_u^2G^T.$$  

In the measurement update step, the Kalman Gain and state vectors are updated by

$$K(k) = P(k|k-1)H^T(HP(k|k-1)H^T + \sigma_v^2)^{-1},$$

$$\hat{x}(k) = \hat{x}(k|k-1) + K(k)(y(k) - H\hat{x}(k|k-1)),$$

$$P(k|k) = (I - K(k)H)P(k|k-1),$$

where $I$ denotes the identity matrix.
3. EVALUATION AND DISCUSSION OF PERFORMANCE

In this section we evaluate the objective performance of the proposed method in terms of segmental SNR and PESQ with comparison to some existing approaches.

3.1. Simulation conditions

Clean female and male speech samples with the sampling rate of 16kHz are generated from TSP database [11]. Non-stationary noise and babble noise sequence taken from NOISEX-92 [12] are added to the clean speech signal with different input SNRs ranging from $-5\text{dB}$ to $10\text{dB}$. The frame size is 512 samples and a 3-level WP decomposition tree with Daubechies1 wavelet is applied to the noisy speech. For VAD initialization, we pick the minimum frame feature value from the first $N = 10$ frames. In thresholding values modification part, we set $\tau = 2$. In this case, if subband frame $\text{SNR}_{\text{seg},n}$ is equal to or higher than $10\text{dB}$, the subband frame has relatively more voice portion. Thus, $e^{-5}(e^{-\text{SNR}_{\text{seg}}}/\tau = e^{-5})$ is almost zero and $\hat{T}_n$ is taken as $T_n$ based on (10). We found that the smoothing parameter $\alpha$ and the scaling parameter $\lambda$ perform reliably when they are in the range of $[0.9, ..., 0.95]$. The LPC order considered in this simulation is set to $p = 8$.

3.2. Objective evaluation and comparison

Three existing methods namely, adaptive threshold (AT) [9], iterative Kalman filter (IKF) [4] and subband iterative Kalman filter (S-IKF) [6], are compared with the proposed adaptive threshold iterative Kalman filter (AT-IKF) method. Standard objective metrics, i.e., segmental SNR and PESQ, are applied for performance evaluations. From Fig. 4, it is observed that in non-stationary noise environment the proposed method achieves the same performance as compared with the IKF and S-IKF methods in terms of segmental SNR. However, the experimental results shown in Fig. 5 demonstrate a performance improvement in terms of PESQ. Moreover, as shown in Fig. 6 and Fig. 7, the proposed scheme outperforms the other methods both in terms of segmental SNR and PESQ.

4. CONCLUSION

In this paper, we have proposed a VAD based adaptive WP thresholding scheme with the IKF for speech enhancement. The noisy speech was first decomposed into 8 subbands. Two features have been chosen for the VAD to detect whether the speech frame of each subband is a voiced or noise frame. Based on the VAD results, the threshold was updated for each frame of different subbands, while each frame was adjusted by adaptive thresholding. Through the inverse WP transform, the pre-enhanced whole-band speech has been obtained. The IKF was then used to further enhance the speech. Based on segmental SNR and PESQ evaluations from extensive simulations, it was shown that the proposed method efficiently improved the speech for a wide range of input SNRs and different kinds of noise environments. Compared with conventional subband KF based methods [4] that each subband needs to be processed by the KF based method, the proposed method reduced the computational complexity since only the full band speech is processed by the IKF. Moreover, a pre-enhanced speech, rather than the direct noisy speech, was processed by the IKF, thus increasing the accuracy of LPCs and noise estimation.

5. ACKNOWLEDGEMENT

The authors would like to thank Prof. Benoit Champagne of McGill University, Montreal, Canada for his comments and suggestions on the paper that have helped improve the presentation and quality of our paper.
6. REFERENCES


