SPEECH ENHANCEMENT BASED ON TURBO ITERATION

Hang Dong, Hong Sun

SPEECH ENHANCEMENT BASED ON TURBO ITERATION

Hang Dong and Hong Sun

Signal Processing Lab, School of Electronic Information
Wuhan University, Wuhan, China
donghang115@yahoo.com.cn hongsun@whu.edu.cn

ABSTRACT

A Turbo iterative method for speech enhancement is proposed. The Kalman filter with the voice generation model and the wavelet threshold filter with the short-term spectrum are combined through Turbo iteration. Both filters work in rotation and each one takes some feedback information from the other filter as a priori condition. Our experiment results show that the Turbo iterative algorithm will converge within 10 iterations, and it achieves a good balance between the noise reduction and voice restoration.

Index Terms— Speech enhancement, Turbo iterative, voice distortion

1. INTRODUCTION

Speech quality and intelligibility might significantly deteriorate in the presence of background noise. In particular, speech coders and automatic speech recognition systems, that were designed or trained to act on clean speech signals, might be rendered useless in the presence of background noise. Therefore, it is essential to include the speech enhancement technique for such systems. In speech communication systems, a variety of approaches based on single-microphone have been used, including spectral restoration, Kalman filter, and wavelet method [1-3].

An iterative process is an effective method for speech enhancement when the property of noise is unknown or the parameters of speech model are hardly to estimate. Lim and Oppenheim have suggested an iterative Wiener filter method [4], which operates on short-term segments of the speech signal. The disadvantage of this method is that no proper convergence criteria exist, and after just a few iterations beyond convergence, the quality of the estimated speech signal becomes degraded. To reduce the musical noise, Ogata [5] adopted an iterative algorithm which uses the output signal of the spectral subtraction method as a new input signal, and the noise spectrum is re-estimated at every iterative processing. Hansen [6] suggested a method that introduces constraints to the estimated all-pole speech parameters, so that they retain speech-like properties. This method applies inter- and intra-frame spectral constraints to ensure convergence to reasonable values and hence improves speech quality. As long as belief propagation is considered, those processes may be said as self-iterative, since the feedback loop is used inner a system.

A very different iterative method is presented in the Turbo code, which introduced by Berrou et al. in 1993 [7], The Turbo code is among the most important developments in the field of coding theory for its excellent performance. It uses two encoders to get a couple of orthogonal codes by interleaver. Two decoders work in a turbo way, and the information of decoding is exchanged between the two decoders. After several iteration (3-5 times), it achieves convergence and gets an amazing performance. Turbo iteration method has been introduced in the image processing, and achieved remarkable results [8].

Speech enhancement approaches are often used in complex environments, such as high levels of ambient noise, or lack of model parameters, etc. On the other hand, a variety of different speech enhancement methods based on different speech models are developed over the past several decades, which have their own advantages and different limitations. The principle of Turbo iteration will be introduced to speech enhancement technology in this paper: use two different speech enhancement systems which based on different models respectively to process the noisy speech, and then exchange the information between the two processing systems. We demonstrate the efficiency of this turbo iterative approach.

In section 2**,** we discuss the two speech enhancement approaches we adopted: Kalman filter based on voice generation model and wavelet filter based on short-term spectrum, and analyze their trends when self-iteration is applied. In section 3 we then describe how turbo iterative processing may be applied to estimate the speech with additive noise at very low SNR. And a detailed description for the algorithm is listed. Finally we present experimental results and draw the conclusions in section 4.

2. AUTO-ITERATIVE SPEECH ENHANCEMENT

2.1. Iterated Kalman filter

Let the signal measured by the microphone be given by

$$y(n) = s(n) + d(n), \tag{1}$$

where $s(n)$ is the clean speech signal, and $d(n)$ is the uncorrelated with $s(n)$ additive background noise.

For each frame, the speech signal is assumed to follow an autoregressive (AR) model [9].

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + g u(n) \tag{2}$$

where $u(n)$ is the excitation signal, assumed to be white noise with zero mean and unit variance, g is a gain factor, and α_k are the AR coefficients (p order).

Assuming that the parameters α_k and g are known, we can use the standard Kalman filter, that provides the optimal minimum mean square error (MMSE) estimate of the state vector. In practice, however, these parameters are not available. The Estimate-Maximize (EM) method is applied to estimate these parameters. It iteratively estimates the speech AR model parameters, and applies the Kalman filter at each iteration step. $\hat{\alpha}_k^{(l)}$, $\hat{g}^{(l)}$ and $\hat{s}^{(l)}(n)$ denote the estimates of α_k, g and $s(n)$ after the l th iteration.

To obtain the parameter estimate at iteration $l+1$, the following two-step EM procedure is adopted.

E-step: On the assumption that $\hat{\alpha}_k^{(l)}$ and y are known, we can estimate $s(n)$ according to the maximal a posterior probability. Using the well-known Kalman filter recursion [2], we can get the estimated signal $\hat{s}^{(l)}(n)$. The E-step is followed by the M-step providing the parameter estimates for the next iteration.

M-step: In the case that $\hat{s}^{(l)}(n)$ is known, $\hat{\alpha}_k^{(l+1)}$ and $\hat{g}^{(l+1)}$ can be deduced from the Yule-Walker equations by using Levinson-Durbin algorithm [10].

It can be proved that each iteration step increases the likelihood of the estimate of the parameters, and at last the convergence will approach to a local maximum of the likelihood function.

However, as additional iterations were performed, Individual formants of the speech consistently decreased in bandwidth and shifted in location. And after several iterations, the noise is reduced while lots of detail feature are lost.

2.2. Iterated Wavelet method

Wavelet de-noises as proposed by Donoho and Johnstone [11] in 1992 has been developed for signals with additive noise, and it has been proved to give good de-noising results. The principle is that noise contributes to the majority coefficients but main feature of speech signal contributes to only a few coefficients in the lower bands. Therefore a threshold can be applied to wavelet coefficients to distinguish noise from signal, by setting the smaller coefficients to zero, we can nearly optimally eliminate noise while preserving the important information of original signal.

The process $\boldsymbol{h}$ of threshold wavelet for speech enhancement is summarized as follows

i. Compute the discrete wavelet transform (DWT) of $y(k)$, and get the wavelet coefficient $W(j,k)$, j is the wavelet level.

ii. Process the wavelet coefficients of each level with nonlinear threshold function, and get the estimation $\hat{W}(j,k)$.

iii. Compute the inverse DWT to reconstruct the estimate value $\hat{\boldsymbol{s}}$.

For threshold in the wavelet domain, we use a multilevel soft threshold function that shows the advantages over single threshold function with respect to wavelet coefficients of each level. The threshold function is given by

$$\hat{W}_j = \begin{cases} 0 & |W_j| \le T_j \\ \mathrm{sgn}(W_j)(|W_j| - T_j) & |W_j| > T_j \end{cases}, \quad (3)$$

where $T_j = \sigma\sqrt{2\log N}/\ln(j+1)$, σ is the noise variance, j is the wavelet level.

In applying the multilevel threshold method to speech signal, the unvoiced sound in the speech signal is damaged. Since the unvoiced sound contains lots of noise-like high frequency components, eliminating them in the wavelet domain can cause severe degradation of intelligibility in the reconstructed signal. Therefore, a self-iterative processing based on threshold wavelet can be employed in the noise space to preserve the unvoiced sound.

$$\hat{n} = y - \hat{s}^{(k)} \quad (4)$$

$$\triangle s' = \boldsymbol{h}(\hat{n}) \quad (5)$$

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} + \triangle s' \quad (6)$$

where $\boldsymbol{h}$ denotes the process of threshold wavelet and k denotes the iteration times.

The above self-iterative algorithm, equations (4)-(6), converges usually in just $k = 3-5$ iterations. However, the capability of de-noising remains rather weak, since the undesired noise is extracted with the unvoiced sound in each iteration step. In fact, the statistical characteristics of the residue noisy speech $\hat{n}$ will be far from the true value.

3. TURBO ITERATIVE SPEECH ENHANCEMENT

The self-iterated methods mentioned above can not achieve the best speech enhancement effect. It is difficult to obtain a preferable performance in noise reduction while preserving a low speech distortion. We propose to use Turbo iteration to integrate the benefits of above two methods.

Speech enhancement is a probabilistic inference problem as a decoding problem. Within this scope, we explore the application of turbo principle to speech enhancement. Firstly, two different kinds of models are proposed to describe the disparate characteristics of speech signal respectively: the one is voice generation model, and the other is short-term spectrum model. According to the two signal models, Kalman and wavelet filter are used respectively. Secondly, the information is exchanged between these two independent filters. The result of one filter fed back to the other filter as a prior knowledge, similar to turbo decoding.

Fig. 1 shows the scheme of turbo iterative processing for speech enhancement with additive noise. Kalman filter is used as Filter 1 with a prior knowledge from Filter 2 and with an output of the estimated speech $\hat{s}$. A wavelet-based filter is used as Filter 2 to extract the significant detail features from the residue noisy speech n based on the result of Filter 1, and it gets an output of the estimated detail features s'.

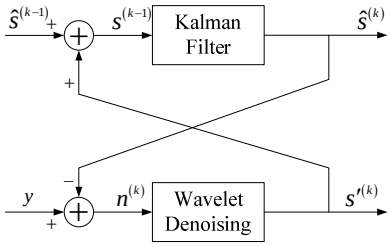


Fig. 1. Scheme of Turbo iterative speech enhancement

Under the scheme of Fig. 1, the algorithm is composed of Kalman filter and wavelet de-noising described above. The procedure of iterative Turbo de-noising is as follows:

0) Initialization ($k=0$): $s^{(0)}=y$, $n^{(0)}=0$

1) Compute the estimated signal $\hat{s}^{(k)}$ by using Kalman filter, from EM method.

2) Get the residue noisy signal $n^{(k)}$ by equation (4).

3) Filter the noisy signal $n^{(k)}$ based on multilevel threshold wavelet by equations (5) and (6) with an output of the detail features $s'^{(k)}$.

4) Plus the detail features $s'^{(k)}$ with $\hat{s}^{(k)}$ as $s^{(k)}=\hat{s}^{(k)}+s'^{(k)}$.

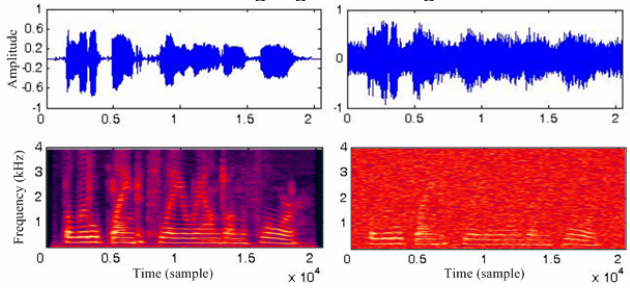
5) $k=k+1$, Go to step 1) if $k<k_{\max}$, or else output $\hat{s}^{(k)}$ as the estimation.

This turbo iterative algorithm propagates the information of Filter 2 (wavelet threshold) to Filter 1 (Kalman filter) as a part of priori knowledge in Kalman algorithm through step 4, and in the other way, it propagates the information of Filter 1 to Filter 2 as a feedback in wavelet algorithm through step 2. This information exchange is of prime importance for the efficiency of Turbo de-noising.

4. EXPERIMENTAL RESULT

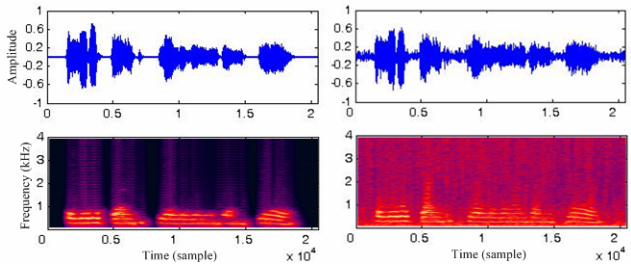
In order to evaluate the performance of the proposed Turbo iteration algorithm, we conduct experiments on 10 speech utterances with three different speakers and 30 sec of speech. The speech is sampled at 8 kHz and quantized to 16 bits. Computer-generated stationary white Gaussian noise is artificially added at 0 dB SNR, its variance is assumed to be perfectly known. A frame size of 32 ms with 50% overlap is used.

A sample noisy utterance is enhanced by using self-iterative Kalman filter, self-iterative wavelet threshold and Turbo iterative method with 10 iterations, respectively. The experiment results show that the self-iterative Kalman filter tends to provide a too suppressed speech signal $\hat{s}^{(k)}$ while some unvoiced speech signal is filtered as noise (Fig. 3 (a)). On the contrary, the self-iterative wavelet threshold tends to extract detail information but a lot of noise rest in the enhanced speech (Fig. 3 (b)). The Turbo iterative algorithm balances well these two trends (Fig. 4). From the example spectrum in Fig. 4, the proposed method can effectively suppress the noise, due to the Turbo iterative scheme. Additionally, it can be seen that the proposed method helps preserve weak speech segment information more than Kalman filter, as shown in the highlighted rectangular areas.



(a) (b)

Fig. 2. (a) Clean utterance in time and spectrogram domains; (b) Noisy utterance in time and spectrogram domains (0 dB SNR AGWN)



(a) (b)

Fig. 3. Enhanced utterance in time and spectrogram domains (a) when the self-iterative Kalman filter is used (10 iterations); (b) when the self-iterative wavelet threshold is used (10 iterations)

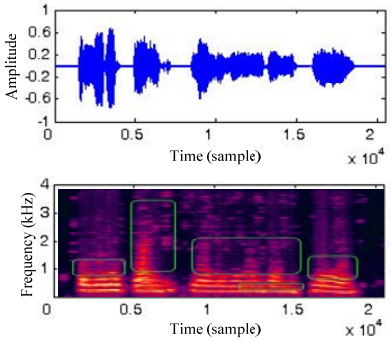


Fig. 4. Enhanced utterance in time and spectrogram domains when the Turbo iterative method is used (10 iterations)

Two kinds of objective tests are conducted: segmental SNR (SSNR) and log spectral distance (LSD). The segmental SNR is computed on segments of 20 ms as

$$SegSNR=\frac{1}{J}\sum_{l=0}^{J-1}10\log_{10}\frac{\sum_{n=0}^{M-1}s^2(n+lM/2)}{\sum_{n=0}^{M-1}\left[s(n+lM/2)-\hat{s}(n+lM/2)\right]^2} \quad (7)$$

The LSD is computed as

$$LSD=\frac{1}{J}\sum_{l=0}^{J-1}\left(\frac{1}{M/2+1}\sum_{k=0}^{M/2}\left[10\log_{10}S_l(k)-10\log_{10}\hat{S}_l(k)\right]^2\right)^{1/2} \quad (8)$$

where the outer summation is a sum over J speech segments, and M is the frame length. $S_l(k)$ and $\hat{S}_l(k)$ are the short-time spectral amplitude of clean speech and the enhanced speech signal of the l th signal segment, respectively.

Table 1 and 2 show the results of the SSNR improvement and the LSD for various iterations respectively, where the Turbo iteration (TI) is compared with the self-iterative Kalman filter (AIKF) and self-iterative wavelet threshold (AIWT). From the results, we see the proposed method is the most advantageous method.

Method	Iterations			
	one	three	five	ten
AIKF	18.76	20.27	19.32	18.92
AIWT	17.32	17.15	16.36	16.21
TI	21.12	21.89	22.03	22.15

Table 1. Comparison of SSNR (in dB) of enhanced signals for various iterations

Method	Iterations			
	one	three	five	ten
AIKF	8.02	7.86	8.04	8.56
AIWT	8.97	8.83	8.78	8.76
TI	7.53	7.42	7.33	7.28

Table 2. Log spectral distortion (in dB) between the clean signal and enhanced signals for various iterations

5. CONCLUSIONS AND FUTURE WORK

In this paper, a novel turbo iterative method for speech enhancement has been presented. We utilize the statistical filter based on voice generation model and wavelet threshold filter based on short-term spectrum to enhance the noisy signal respectively, the two filters work in rotation. The filters extract the parameters of voice model and the information of unvoiced speech from the results of other filter, and then re-processing is executed. With these kind extrinsic references the two filters work in Turbo iterative method so that to improve the performance of each filter. Compared with self-iterative algorithm, our Turbo iterative method takes benefits of noise reducing and detail preserving synchronously. The experimental results show the proposed method will be convergent after 10 iterations. In addition, a preferable filter can be embedded in our method to pursuit the preferable performance in the future.

6. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 3, pp. 113-120, 1979.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech Audio Processing*, vol. 6, no. 4, pp. 373-385, 1998.

[3] N.A. Whitmal, J.C. Rutledge, and J. Cohen, "Wavelet-based noise reduction," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 3003-3006, 1995.

[4] J.S. Lim, A.V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, no. 3, pp. 197-210, 1978

[5] O. Shinya, and S. Tetsuya, "Reinforced Spectral Subtraction Method to Enhance Speech Signal," *Electrical and Electronic Technology, TENCON Proceedings of IEEE Region 10 International Conference on*, pp. 242-245, 2001.

[6] J.H.L. Hansen, and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795-805, 1991.

[7] C. Berrou, A. Glavieux, and P. Ritimajshima, "Near Shannon-limit error-correcting coding and decoding : turbo codes," *IEEE International Conference on Technical Program, Conference Record*, vol. 2, pp. 1064-1070, 1993.

[8] H. Sun, H. Maitre, and B. Guan, "Turbo image restoration," *Proceedings of Seventh International Symposium on Signal Processing and Its Applications*, pp. 417-420, 2003.

[9] J. Lim, and A.V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197-210, 1978.

[10] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, 1996.

[11] D.L. Donoho, "De-noising by soft-thresholding," *IEEE Transaction on Information Theory*, vol. 41, no. 3, pp. 613-627, 1995.