interspeech 2008 incorporating SST 08

A Wavelet Based Speech Enhancement Method Using Noise Classification and Shaping

Mahdi Mohammadi[1], *Behzad Zamani*[1], *Babak Nasersharif*[2], *Mohsen Rahmani*[1], *Ahmad Akbari*[1]
[1] Iran University of Science and Technology, Tehran, Iran
[2]Computer Engineering Department, Faculty of Engineering, University of Guilan
{mh_mohammadi, bzamani, m_rahmani, akbari}@iust.ac.ir, nasersharif@guilan.ac.ir

Abstract

Speech enhancement systems performing in Fourier or wavelet domain usually generate musical noise and distortion. It is possible to reduce musical noise and speech distortion by shaping residual noise. In this paper, we propose to implement a noise shaping method for a wavelet based noise reduction system. For noise shaping, we propose a noise classification method based on spectral shape of input noise. Using this classification method, we transform input noise to another noise which is more acceptable from listening point of view. Objective and subjective test results show that using noise shaping method; we obtain less distortion in speech signal in comparison to basic wavelet noise reduction system. Furthermore, listening test results illustrate that background shaped noise is less annoying.

Index Terms: speech enhancement, noise shaping, noise classification

1. Introduction

Speech enhancement methods have been widely used in noisy environments for many different applications such as mobile phone systems, speech coding systems and hearing aids devices [1,10,11] . Frequency based speech enhancement methods modify the frequency spectrum of noisy signal to reduce residual noise and speech artifacts in enhanced signal[4]. These methods create distortion and musical noise in enhanced signal due to their limited time-frequency resolutions. This problem can be solved using wavelet transform.

In Speech enhancement methods using wavelet transform, we remove noise components by thresholding the wavelet coefficients[3]. Wavelet coefficients are compared to a threshold and it is determined which coefficients must be set to zero. The proper value of threshold can be determined in many ways. In [1], Donoho proposed a thresholding method that performs well for white noise, while it has not good performance in presence of color noise. In addition, using threshold methods sometimes lead to removing high frequency components of speech signal and so decreasing quality of speech. There are two approaches in order to overcome this weakness. In first solution, some researchers proposed to improve thresholding method or threshold values [2]. On the contrary, in the second solution, some researchers use Wiener filter in wavelet domain [9] .In this approach, filter parameters are adjusted based on statistical condition of wavelet coefficients.

In this paper, we use Wiener filter in wavelet domain for speech enhancement. We also adjust Wiener filter parameter in such way that does not remove noise completely in wavelet domain. This can reduce the distortion generated by removing all of noise. So, we remain some background noise by adjusting Wiener filter. Furthermore, we shape this remaining noise to a desired noise for listeners. This desired noise should be similar to original noise and should reduce the distortion made by noise shaping method. For this purpose, we classify input original based on their spectral shape and determine a desired noise for each class. In this way, we can choose a desired noise which its spectral shape is similar to input noise spectral shape. This results in less distortion in speech signal. The general proposed method block diagram is shown in Figure (1).

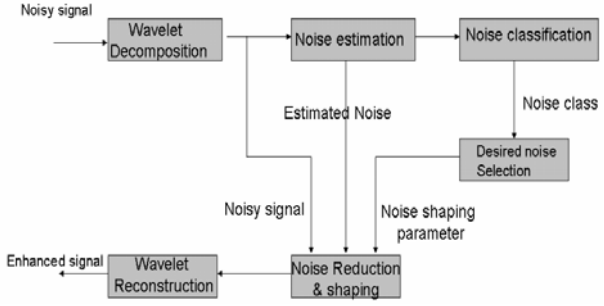

Figure 1: *The general proposed method block diagram*

The remainder of this paper is organized as follows. In Section 2, we explain the noise reduction method in wavelet domain using Wiener filter. In Section 3, we describe our noise shaping method. In section 4, we define our noise classification method in wavelet domain. Section 5, we give detail of desired noise selection. In section 6, we give details of our noise reduction system. Section 7 includes our experiments and results. Finally, we give our conclusion in Section 8.

2. Noise reduction based on wiener filter in wavelet domain

In wavelet based speech enhancement systems, input signal is decomposed into some sub-bands in decomposition phase and power of noise is decreased with applying noise reduction filter to each sub-band. After that, in reconstruction phase, enhanced signal is achieved by applying wavelet reconstruction filters to enhanced subbands. Wavelet packet is a type of wavelet transform which is employed in this paper [5]. In this paper number of subbands in the wavelet packet tree is 16. In speech enhancement systems based on wavelet transform, power of each subband is considered as indicator of subband and applied in noise reduction filter. General equation of enhancement filter in wavelet domain is shown in equation (1).

$$\hat{S}(m,f,j) = H(m,f) \times X(m,f,j) \quad (1)$$

In this equation, X(m,f,j) is j th coefficient of f th subband in m th frame of noisy signal. $\hat{S}$(m,f,j) is j th coefficient of f th subband in m th frame of enhanced signal and H(m,f) is calculated in f th subband.

Wiener filter is a known filter in speech enhancement applications applied in frequency and wavelet domains [6]. Wiener filter in wavelet domain is computed as in (2).

$$H(m,f) = \frac{\sigma_{s,f}^2}{\sigma_{s,f}^2 + \sigma_{n,f}^2} \quad (2)$$

In this equation, $\sigma^2_{s,f}$ is power of clean signal in f th subband and $\sigma^2_{n,f}$ is power of input signal in f th subband. But in real-time applications $\sigma^2_{n,f}$ is not available. So an estimation of $\sigma^2_{n,f}$ is used in Wiener filter equation. Using this estimation, we rewrite equation 2 as:

$$\overline{\sigma}^2_{S,f} = \sigma^2_{X,f} - \overline{\sigma}^2_{n,f}$$
$$H(m,f) = \frac{\overline{\sigma}^2_{S,f}}{\overline{\sigma}^2_{S,f} + \overline{\sigma}^2_{n,f}} \quad (3)$$

With adding term $\alpha\overline{\sigma_{n,f}}$ to the numerator of Wiener filter, we obtain equation (4). This term is used to control the amount of noise reduction. In [7] authors showed that using this term, amount of speech distortion is reduced in enhanced signal.

$$H(m,f) = \frac{\overline{\sigma}^2_{S,f} + \alpha\overline{\sigma}^2_{n,f}}{\overline{\sigma}^2_{S,f} + \overline{\sigma}^2_{n,f}} \quad (4)$$

Controlling amount of noise reduction causes that some noise remain in the background. But, sometimes remained noise is annoying for listeners. In this paper, we shape the remained noise to another noise and we try to solve the mentioned problem. This noise shaping process is described in the next section.

3. Noise Shaping

Noise shaping can be defined as a process which transforms energy of input noise subbands to the corresponding subbands of a desired noise [8]. For this transformation, a subband dependent parameter is should be determined. We call it as *noise shaping parameter* or α(S) which S is subband dependent variable. Based on this parameter, input noise is shaped to a desired noise. We can describe the noise shaping idea as following. If energy of input noise in subband i, is greater than energy of desired noise in subband i, energy of input noise in subband i is reduced to corresponding subband of desired noise. When energy of input noise in subband i is lower than corresponding subband of desired noise, the input noise subband don't change. Equation (5) shows the overall process for calculating α(S). In this equation $N_D(S)$ is energy of S^{th} subband of desired noise and $N_I(S)$ energy of S^{th} subband of input noise.

$$\alpha_S(S) = MIN\left(\frac{N_D(S)}{N_I(S)}, 1\right) \quad (5)$$

After calculating α(S), it applies to energy of input noise subbands. Based on this shaping parameter, power of shaped noise for each subband is computed based on equation (6).

$$N_R(S) = \alpha_S(S) \times N_I(S) \quad (6)$$

4. Proposal of a noise classification method

A proper desired noise should have some important properties. One of these properties is that the selected desired noise must have similar spectrum to the input noise. This aspect guides us to propose a noise classification method. Based on this noise classification method, we can define a desired noise for each class. Spectrum of desired noise for each subband should be similar to the spectrum in its class. Our proposed noise classification method classifies input noise based on its spectral shape. We define 4 classes of noise signals: low-pass, high-pass, band-pass and white. Spectral shape is defined in frequency domain, so we must define a new edition of spectrum in wavelet domain. For this purpose we calculate energy of each subband in decibel and interpret this energy of subbands as spectrum in wavelet domain. We name this as power distribution of subbands. Figure (2) shows power distribution of subbands for a highpass noise in wavelet domain.

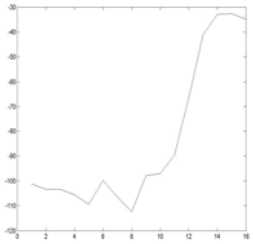

Figure 2: *Power distribution of subbands for a highpass noise*

Main idea in classification method is based on estimation of power distribution with a linear curve (ax_i+b). For this purpose, we should solve Σ (y_i-(ax_i+b))=0, so *a* can be achieved based on equation (7). We call this feature as slope of power distribution.

$$a = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (7)$$

In this equation, n is number of subbands, y_i is power of i^{th} subband and x_i is index of subband (x_i= {1, 2…, number of subband}). Computed slope is positive for high-pass class and is negative for low-pass class. Therefore, we define 3 rules for our classification method:

- If (a<-1), input noise is lowpass
- If (a>1), input noise is highpass
- Otherwise input noise is white or bandpass.

Slope feature is not enough to separate between band-pass and white classes. So, we introduce another feature besides slope. We name this feature as GTA (number of samples Greater Than Average).

For calculating this feature, we use the average of power distribution for input noise. In white noise, number of subbands greater than average, are about 50% of all subbands. But, in band-pass noise, number of subbands greater than average is smaller than 50% of all subbands. Based on this feature, we can distinguish between band-pass noise and white noise. Figure (3) shows two examples for slope and GTA features. As shown in figure (3), grey curve shows the power distribution of subbands and black line shows the estimation of power distribution with a linear curve. And this figure shows the number of samples greater than average for a bandpass noise.

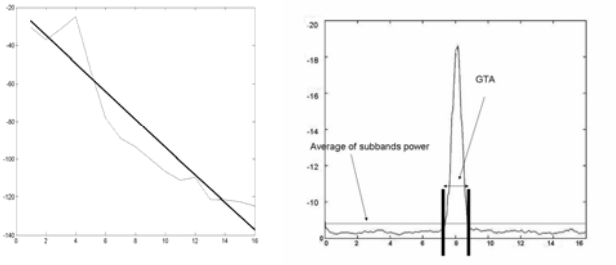

Figure 3: *Two examples for slope and GTA features*

In order to evaluate our noise classification methods, we constructed a noise database. Noise database includes: 48 low-pass noises with 3 different power levels (20 dB, 30dB and 40 dB) with different cut-off frequencies, 48 high-pass noises with the same 3 levels and different cut-off frequencies, 56 band-pass noises with different central

frequency and the same 3 mentioned levels, 6 white noises with 6 different levels. Finally some 42 other noise samples were added to our database. These samples include some real noises like pink, car and F16 noises and some synthetic noises with selected power spectral shapes. Correct recognition rates for studied noise classification methods are shown in table (1).

As shown in the table, overall accuracy of proposed noise classification method is acceptable on 200 different noises.

Table 1.

Noise Type	Classification Rate	Noise Type	Classification Rate
HighPass	91.6%	BandPass	86%
LowPass	91.6%	White	100%
Other noises	90%		
Total 90%			

5. Desired noise selection

As mentioned, we must select a desired noise for each class to transform input noise to it. Desired noise for each class should have two important properties. 1). similar spectrum to input noise spectrum. 2). More acceptable from listeners view point. We select a desired noise for each class based on listening test. For this purpose, we generate some different noises for each class and then listener select a less annoying noise for each class. All of these generated noises have the same energy and frequency sample rate 8 kHz. Therefore, listeners select desired noises in fair conditions. Selected noises for each class are as follows.

Lowpass class: with generating different lowpass noises, listeners select car noise as the desired noise for this class.

White class: all the noises placed in white class, have the same spectrum. If we select a white noise as desired noise for this class, then we should shape spectrum of a white noise to another white noise. So, this process is not shaping. For white class, we select a lowpass noise with a spectrum very similar to car noise, as desired noise. We select this desired noise based on 2 reasons. At first, lowpass noises are less annoying to listeners. Second, most of speech power is concentrated in low subbands. Consequently, with shaping a white noise to a lowpass noise, amount of speech artifact is reduced. This is due to that noise power in low subbands, have little changes based on this shaping process.

Highpass class: it was a difficult task to select a highpass noise as desired noise, for listeners, because highpass noises are annoying for hearing. For this class, we generate more highpass noises and more listeners listen to them. Finally, listeners select a highpass noise which most of its energy is concentrated in subbands higher than 3500 Hz as desired noise of this class.

6. Noise reduction and shaping

For shaping input noise to a desired noise, we change equation (4). In this new equation, amount of noise reduction in each subband is changed. Then, we use α(f) in shaping and reduction filter as in equation (8).

$$H(m,f)=\frac{\overline{\sigma}_{S,f}^2+\alpha(f)\overline{\sigma}_{n,f}^2}{\overline{\sigma}_{S,f}^2+\overline{\sigma}_{n,f}^2} \quad (8)$$

The main diagram of shaping process is summarized in Figure (4). As shown in figure (4), amount of noise is estimated in each subband. After that, with applying noise classification to estimated noise, class of input noise in current frame is determined. Based on this determined class, we calculate noise shaping parameter α(f) (equation (8)). With using α(f) and equation (8), input noise in current frame is shaped to the desired noise.

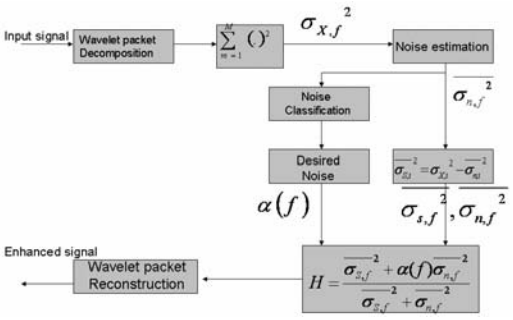

Figure 4- Noise reduction and shaping diagram

7. Experiments and results

In our experiments, we used 12 utterances from TIMIT dataset for evaluating performance of proposed shaping method which contain 6 male and 6 female. Length of each utterance is 3 second with frequency sampling rate 8 kHz. 4 types of noise are added to these signals with 4 SNR 0, 5, 10, 20 dB (white, car, babble and a highpass noise). So we generate 12×4×4 (192) different signals for evaluating our method. Our experiments contain 3 objective and 1 subjective tests. In objective tests, we used spectrum distance, cepstrum distance and SNR improvement (SNRI) as our measurements. Subjective test contains two listening test. In the rest of paper, we describe our experiments results.

7.1. SNR Improvement (SNRI)

The first objective measure for evaluating proposed method is SNRI. Amount of SNRI for each type of noise is the average SNRI on 12 enhanced signals. Figure (5) shows SNRI measure for white, car, babble and highpass noises for different input SNRs. In case of car noise, input noise and desired noise are the same, so amount of SNRI in this condition is not noticeable. In presence of white noise, amount of SNRI is noticeable because input noise is shaped to a lowpass noise.

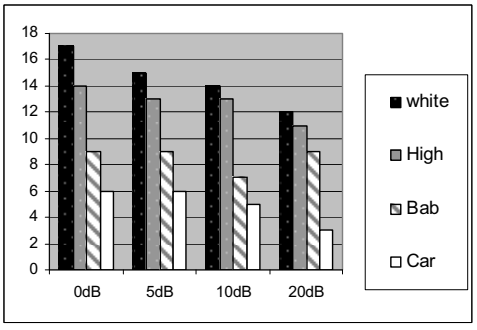

Figure 5: *SNR improvement measure*

7.2. Cepstral distance

The second criterion used for measuring amount of speech artifact, is cepstral distance (amount of noise reduction with shaping and without shaping are the same). Equation (10) shows this criterion. In this equation, M is number of frames; L is number of cepstral coefficient, $C_c(i,j)$ is j^{th} cepstrum coefficient of i^{th} frame of clean signal. $C_E(i,j)$ is j^{th} cepstrum coefficient of i^{th} frame of enhanced signal (with shaping or without shaping).

$$Dis\_ceps=\frac{\sum_{i=1}^{M}\sqrt{\sum_{j=1}^{L}\left(C_C(i,j)-C_E(i,j)\right)^2}}{M} \quad (10)$$

As illustrated in figure (6), cepstral distance for shaping method is less than without shaping method, especially in

input SNR 0 and 5 dB. It shows that cepstral coefficient of shaped enhanced signal is closer to cepstral coefficient of clean signal in comparison with non-shaped enhanced signal. So amount of speech artifact with shaping method is less than method that doesn't use shaping.

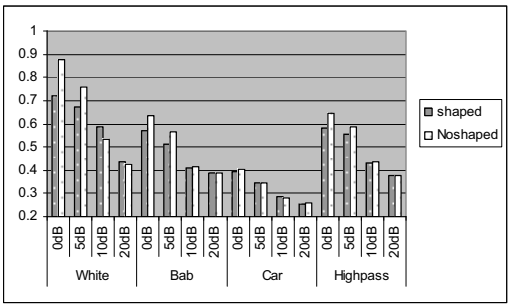

Figure 6: *Cepstral distance measure*

7.3. Listening test

We use two different listening tests as subjective measure. These tests are two preference listening tests which listener listens to 2 enhanced signals (with shaping and without shaping methods) and gives scores to both of them. (Sum of these two scores should be 100). If assigned number for each signal is greater than the other, then it shows that listener prefers this signal to the other. Difference between these two scores shows the difference between qualities of these 2 signals.

In the first listening test, we have some pure noise parts between speech segments. In this listening test, listener gives score to the enhanced signals based on two aspects, remained background noise and acceptability of remained noise. Figure (7) shows the results of these experiments for white, car, babble and highpass noises. Each value in this figure is average of listening test scores for 4 enhanced signals in input SNR values of 0, 5, 10 and 20 dB.

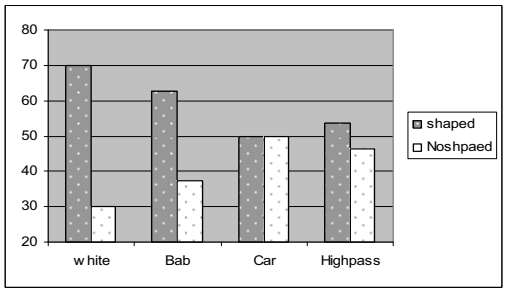

Figure 7: *Listening test for background noise*

In the second listening test, speech segments are selected. After that, listeners give score to them. In the other word, there isn't any pure noise segment in this experiment. The main idea of this process is that we can compare enhanced signals from view point of speech artifact. Figure (8) shows results of this experiment. As shown in this figure, in presence of white and babble noises, 70% of listeners prefer shaped noise. But in presence of car and highpass noises, this value is about 50% of listeners. Because in case of car noise, input noise and desired noise are the same and in case of highpass noise, selection of desired noise is very difficult task and most of these noises are annoying.

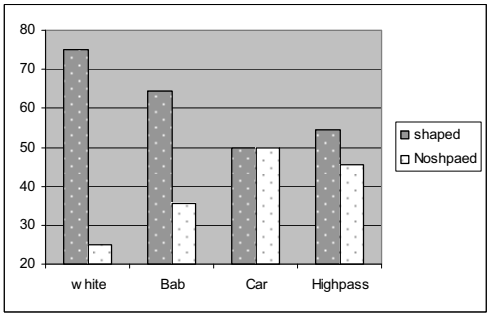

Figure 8: *Listening test for speech artifact*

8. Conclusion

In this paper, we propose a noise reduction method in wavelet domain. In this method, a noise shaping algorithm is proposed which improves the performance of noise reduction system. This noise shaping algorithm uses a noise classification method in wavelet domain. In this classification method, we have 4 classes: white, lowpass, highpass and bandpass noises. For each class of noise, we select a desired noise. Then, we shape input noise to determined desire noise using shaping process. The experimental results shows that proposed noise shaping method can reduce speech artifact, successfully. In listening tests, 70% of listeners prefer noise reduction with shaping method to noise reduction without shaping method. Objective measures show acceptable performance of proposed method, especially in presence of white and babble noises.

9. References

[1] Donoho, D., "De-noising by soft-thresholding," in *Proc. IEEE Trans. Inform. Theory 41*, 1995, pp. 613–627.

[2] McAulay, M., "Speech enhancement using a soft-decision noise suppression filter," in Proc. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, April, 1980, pp. 137–145,.

[3] Yann, I.و Ngee, S. Kiat, C. "Wavelet for speech denoising". *Proc IEEE TENCON 97*, vol. 2, pp. 479-482.

[4] Ephraim, Y. وMalah, D.و "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, " in *Proc IEEE Trans. on Acoustics, Speech, and Signal Processing*, Dec. 1984,Vol 32, pp.1109-1121.

[5] Addsion, P. S. *The Illustrated Wavelet Transform Handbook*, Institute of Physics, ISBN 0-7503-0692-0, 2002.

[6] Wiener, N., *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley, 1949

[7] Rahmani , M., Mohammadi, M., Akbari A., "Background noise control for speech enhancement", in *Proc.14th ICEE 2005.*

[8] R. Wannamaker, "Psychoacoustically Optimal Noise Shaping", in *Proc, J. Audio Eng. Soc. 40 (7/8)*, 1992, pp.611-620.

[9] Asefa, D. , Mital, D., Haque S., "Restoration of fMRI Signal using Wiener Filter in a Wavelet Domain", *The Internet Journal of Medical Informatics* 2005.

[10] Herman, D., Brennan, R., sheikhzadeh, H., Cornu, E., "Low Power Implementation of The Bluetooth Subband Audio Codec," in Proc. *International Conference on Acoustics Speech and Signal Processing, Vol. 5, pp.365-368,* 2004.

[11] Ching L., Hsiao, W., "Enhancement of single channel speech based on masking property and wavelet transform,". *Speech Communication*, Volume 41, Issues 2-3, pp. 409-427 , October 2003