2D Spectrogram Filter for Single Channel Speech Enhancement

HUIJUN DING, ING YANN SOON, CHAI KIAT YEO*, and SOO NEE KOH
School of Electrical and Electronic Engineering
*School of Computer Engineering
Nanyang Technological University
S2.2-B4-02, Media Tech Lab, 50 Nanyang Avenue, Singapore 639798
SINGAPORE

Abstract: In this paper, we propose a novel approach for single channel speech enhancement by exploiting the correlation among 2D transform coefficients, which has been previously neglected by traditional speech enhancement methods. Our approach makes use of a time-frequency representation (spectrogram) of the input signal and a novel 2D spectrogram filter (2DSF)is designed to provide a good estimate of the original clean speech. The 2DSF, which is easy to implement, comprises a hybrid Wiener filter, statistical classification and a postprocessor. The efficiency of our proposed approach is proven via both objective and subjective evaluations.

Key–Words: Speech enhancement, Wiener filter, noise reduction, spectrogram filtering

1 Introduction

Speech enhancement is not just a research area for academic but is also readily exploited in industrial applications. As we cannot avoid the situation of speaking in a noisy environment in the real world, speech enhancement technique is needed to eliminate the annoying background noises. The purpose of speech enhancement is to ensure speech intelligibility and alleviate listening fatigue. The problem is made tougher in the case of single channel speech and especially when high quality is required even under a low input signal-to-noise ratio (SNR). Many researchers are thus devoted to work on this problem in order to improve the performance of the enhancement techniques.

Generally, spectral subtractive algorithm is simple and sufficient in applications which require lower quality enhanced speech outputs. Wiener filter is another commonly used estimation method to achieve resultant speech with minimum mean square error. Both these methods operate in the frequency domain and are relatively efficient especially in noise removal in the noise-only period given that speech is not continuously present at all times. However, these two conventional methods and other traditional techniques assume there is no relationship between the different frequency coefficients. Research results from Evans [1] show that correlation exists among different time frames. By adopting image processing techniques, Evans has applied the morphological filter, the opening operator with erosion and dilation, to the 2D time-frequency arrangement of the input speech signal [1]. This purely 2D processing algorithm achieves good result despite the fact that it does not exploit the specific features of speech spectrogram. According to the characteristics of speech spectrogram, Goh [2] proposed a spectrogram filtering algorithm, so does Z. Lin [3], for better performance. But these methods tend to be postprocessors which must follow some classic noise reduction techniques, such as MMSE or spectral subtraction.

As for our proposed approach, we adopt a more unified 2D spectrogram filter based on the work of Soon [4] and Goh [2]to solve the spectral filtering problem. First, the hybrid Wiener filter [4], consisting of 1D Wiener filter and 2D Wiener filter, is used to remove the noises from the spectrogram. Next a more efficient statistical classification technique is adopted to distinguish speech component from non-speech component, the latter is marked for further processing. Finally a blade postprocessor [2] maintains those low-energy speech components which have been erroneously classified as non-speech components and further suppress isolated musical tones. The proposed approach is evaluated by objective measure and informal subjective measure which show that it is superior to both the former algorithms.

2 Methodology

First, short-time Fourier transform is performed on the noisy speech to facilitate the time-frequency analysis and the generated spectrogram is filtered by the hy-

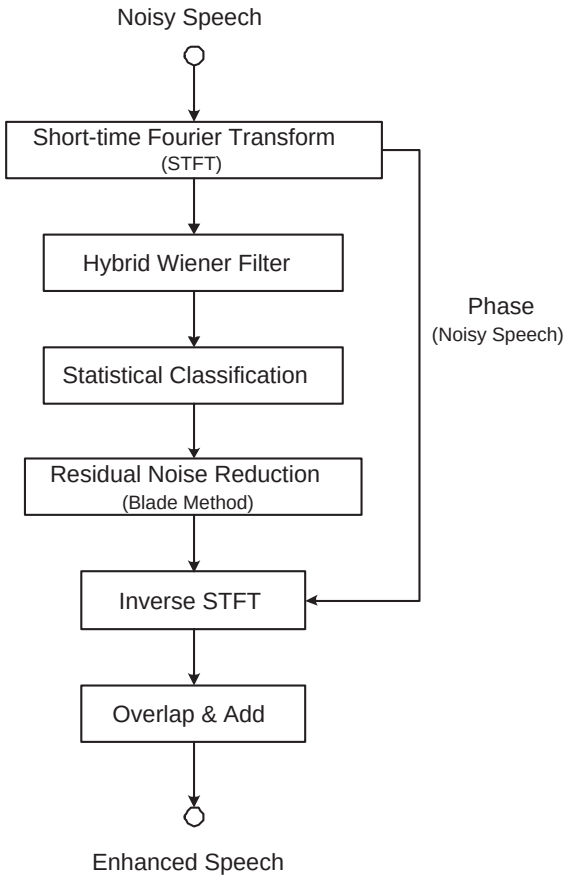

Figure 1: *Block diagram of the 2DSF Algorithm*

brid Wiener filter. We then utilize a statistical classification approach to differentiate the speech component from the non-speech component to improve the algorithm's efficiency before a postprocessor is applied. Fig. 1 shows the overall block diagram of the 2DSF approach which is very simple and easy to implement.

2.1 Short-time Fourier Transform

The clean speech signal $x(m)$ is corrupted by independent additive noise signal $n(m)$, resulting in the noisy speech signal $y(m)$, described as follows in the time domain

$$y(m) = x(m) + n(m) \tag{1}$$

The spectrogram used in our time-frequency analysis is obtained from the short-time Fourier transform of noisy speech. The input signal is multiplied by a window function, such as the Hanning window with a frame size of 256 and with 75% overlap. Thus, the one-dimensional Fourier transform of each window forms a column vector of the 2D matrix with the horizontal axis representing the time and the vertical axis representing the frequency. The intensity of each point in the 2D representation is the speech energy at the given time and frequency. The noisy speech can be expressed in the time-frequency domain as

$$Y(u,v) = X(u,v) + N(u,v) \tag{2}$$

where u is the time frame index and v is the frequency index. $Y(u,v)$, $X(u,v)$ and $N(u,v)$ indicate the complex frequency expressions of the noisy speech, clean speech and noise signal, respectively. As we all know, the complex number $Y(u,v)$ can be expressed as a product of magnitude and phase as shown in Eq. (3). The magnitude of this 2D representation maps to the spectrogram, and the phase information should be preserved for future operation.

$$Y(u,v) = |Y(u,v)| \exp\left(j\phi_Y(u,v)\right) \tag{3}$$

2.2 Hybrid Wiener Filter

The noise is first filtered off by the hybrid Wiener filter from the spectrogram. The reason for calling the filter hybrid is that it consists of two types of Wiener filter, namely, 1D and 2D. The combination of these two parallel channels lead to an impressive result.

The popular 1D Wiener filter [5] is based on the a-priori SNR $\xi(u,v)$ which satisfies

$$W_{1D}(u,v) = \frac{\xi(u,v)}{\xi(u,v)+1} \tag{4}$$

The a-priori SNR can be estimated by the well-known decision-directed approach, explained in [6]. The enhanced spectrogram is estimated from the noisy speech magnitude by the 1D Wiener filter:

$$|\hat{X}_{1D}(u,v)| = W_{1D}(u,v)|Y(u,v)| \tag{5}$$

The 2D Wiener filter exploits the 2D correlations independently without attenuating the coefficients in contrast to some traditional noise reduction techniques. For the purpose of explaining the algorithm, a 2D noise model should be introduced first. An AC component and a DC component comprise this noise model:

$$|N(u,v)| = N_{AC}(u,v) + N_{DC}(u,v) \tag{6}$$

where $N_{DC}(u,v) = E[|N(u,v)|]$, $E[.]$ denotes the expectation function. The following local information at (u,v) is also needed for the 2D Wiener filter:

- Local mean $\overline{|Y(u,v)|}$
- Local variance of noisy speech $\sigma_Y(u,v)^2$

- local variance of AC noise component $\sigma_{N_{AC}}(u,v)^2$

All these values can be computed within a small region at the center of (u,v), for example, a scope of 3 by 3 square. The 2D Wiener filtered speech [4] can thus be described as follows:

$$\begin{aligned} |\hat{X}_{2D}(u,v)| &= \frac{\sigma_Y(u,v)^2 - \sigma_{N_{AC}}(u,v)^2}{\sigma_Y(u,v)^2} \\ &\times \left(|Y(u,v)| - \overline{|Y(u,v)|}\right) \\ &+ \overline{|Y(u,v)|} - N_{DC}(u,v) \end{aligned} \quad (7)$$

If further noise suppression is required, an iterative step can be implemented in the 2D Wiener filter to repeatedly reduce the noise for the desired situation.

The hybrid Wiener filter just simply extracts the minimum value of the 1D and 2D Wiener filter as the output spectrogram magnitude:

$$\hat{X}(u,v) = \min\left(|\hat{X}_{1D}(u,v)|, |\hat{X}_{2D}(u,v)|\right) \quad (8)$$

2.3 Statistical Classification

Although the residual noise is low after the hybrid Wiener filtering, the resulting spectrogram still needs further processing by the postprocessor. A speech/non-speech classification should be implemented first for the postprocessor according to [2]. The mentioned method works on the histogram of the spectrogram, which is too time-consuming and difficult for the classification decision to be made when the discriminating feature is not obvious. Therefore, a more efficient statistical approach is adopted here to benefit the postprocessing process. The classified speech component will be preserved and the non-speech component will be filtered by the postprocessor.

In this paper, this statistical classification approach is proposed as an estimator of probability of speech absence. Reference [7] provides the detail derivations. Considering the two classes, non-speech component H_0 and speech component H_1, the maximum likelihood of obtaining the magnitude of a noisy speech is shown in Eq. (9) and Eq. (10). The magnitudes of noisy speech $|Y(u,v)|$, clean speech $|X(u,v)|$ and noise signal $|N(u,v)|$ are simply denoted as A_Y, A_X and A_N respectively.

$$P(A_Y|H_0) = \frac{2A_Y}{E[A_N^2]} \exp\left(\frac{A_Y^2}{E[A_N^2]}\right) \quad (9)$$

and

$$\begin{aligned} P(A_Y|H_1) &= \tfrac{2A_Y}{E[A_N^2]} \exp\left(-\tfrac{A_Y^2+A_X^2}{E[A_N^2]}\right) \\ &\times I_0\left(\tfrac{2A_Y A_X}{E[A_N^2]}\right) \end{aligned} \quad (10)$$

I_0 indicates the zero order modified Bessel function. In the case of $P(Y(u,v)|H_1) > P(Y(u,v)|H_0)$, the point of consideration is either classified as the speech component marked with 1, or as the non-speech component marked with 0. The discriminating function is shown in Eq. (11).

$$\exp\left(-\hat{\xi}(u,v)\right) I_0\left(2\sqrt{\frac{A_y^2}{E[A_N^2]}\hat{\xi}(u,v)}\right) > 1 \quad (11)$$

The outcome of this process is a binary spectrogram. As the residual noise will be further suppressed, a technique, such as median filter, should be applied to the binary spectrogram to get an over-suppressed spectrogram, which confirms the speech presence class. Meanwhile, the speech absence class is not exactly accurate, since some speech components with low energy may have been erroneously classified as noise. It is intuitive to therefore apply a postprocessor on the speech absence class to reduce the annoying residual noise, while let the misclassified speech components untouched.

2.4 Blade Postprocessor

As mentioned above, the blade postprocessor aims to further process the non-speech component while retaining the speech component. First of all, we should observe the shapes of the residual noise and speech components when choosing a suitable postprocessor. Clear distinctions, namely, isolated peaks and short ridges in contrast with long ridges paralleled with valleys, are observed from the filtered spectrogram.

The postprocessing method is proposed in [2] to attenuate the residual musical noise, especially for the short, ridge-like noise which poses an obstacle for some other time-frequency postprocessors such as the one in [5]. The main contribution of this blade postprocessor is the definition of 16 blades which are actually the 16 orientations with respect to a given point (referred to as the central point). These blades are illustrated in Fig. 2. Each blade consists of 7 points, inclusive of the central point, and is used for the computation of variance. The blade with the minimum variance among the 16 values is selected to represent the class attributes of the central point. This means the median value of this blade is used to replace the original value of the central point, if the median value is not larger than the existing magnitude at the central point. Otherwise, the value of the central point will not be altered. The postprocessed spectrogram is thus as follows.

$$\hat{X}(u,v) = \min\left(\hat{X}(u,v), \underset{(m,n)\in\vec{B}_{\min(var)}}{median}\left(\hat{X}(m,n)\right)\right) \quad (12)$$

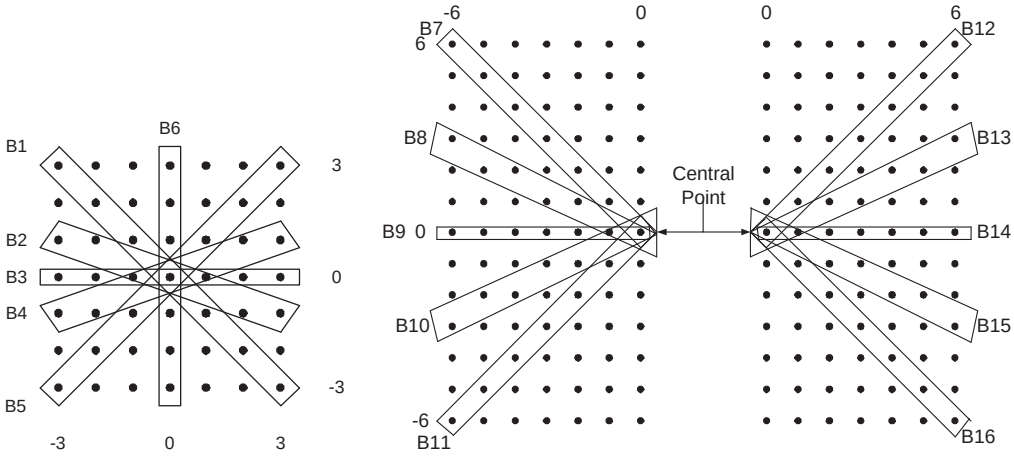

Figure 2: Sixteen blades of the different directions

where $\vec{B}$ represents a blade of a central point, and $\vec{B}_{\min(var)}$ is that certain blade with minimum variance among the 16 blades.

By the selective replacement of the spectrogram magnitude with the median value of the minimum-variance blade, isolated noise points are effectively suppressed; while the speech points which are previously misclassified are unchanged or only slightly altered to preserve the speech intelligibility instead of being erroneously removed as noise.

3 Evaluation and Discussion

Ten speeches from the TIMIT database are tested here. They are resampled at 8 kHz, quantized at 16 bits, and corrupted by various types of noise from the NOISEX database. The noises range from 0 dB to 10 dB and include white noise, car noise, fan noise and f16 aircraft noise.

Two former methods, hybrid Wiener filter (HWF) based on the 2D Fourier transform in [4] and spectral subtraction with a blade postprocessor (BP) in [2], are compared with the proposed 2DSF. The Perceptual Evaluation of Speech Quality (PESQ) measure, which aims at objectively predicting the subjective Mean Opinion Score (MOS) in an ITU-P.800 listening setup, is used in the experiments. It has been shown that the PESQ is more reliable and more correlated with MOS than some traditional measures, such as SNR and Segmental SNR [8].

The results of the comparison are illustrated in Fig. 3 to Fig. 6. The 2DSF shows a significant improvement compared to the others. This observation confirms the efficiency and worthiness of the proposed approach.

Some informal subjective listening tests are also carried out. The same conclusion that the proposed method 2DSF outperforms the other two former work further vindicates the results of objective measure.

4 Conclusion

The problem of 2D spectrogram filtering is discussed here. The new approach proposed substantially suppresses the noise on the spectrogram. Its hybrid Wiener filter exploits not only the 1D property but also the 2D correlation to better preserve speech intelligibility while effectively suppressing the noise. Moreover, the speech content is further preserved by the blade postprocessor which has proven itself to be an effective post processing tool. The lower complexity of computation, which is largely due to the use of statistical classification has simplified the implementation. Objective measure, PESQ, together with informal subjective measure vindicate the superiority of the 2DSF approach.

References:

[1] N. W. D. Evans, J. S. Mason, and M. J. Roach, "Noise Compensation using Spectrogram Morphological Filtering," in *ProcEEDINGS 4th IASTED International Con. on Signal and Image Processing*, 2002, pp. 157–161.

[2] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 287–292, 1998.

[3] Z. Lin and R. Goubran, "Musical noise reduction in speech using two-dimensional spectrogram en-

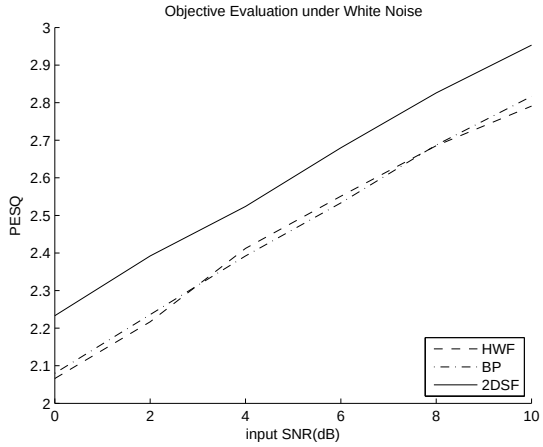

Figure 3: *PESQ measure under white noise*

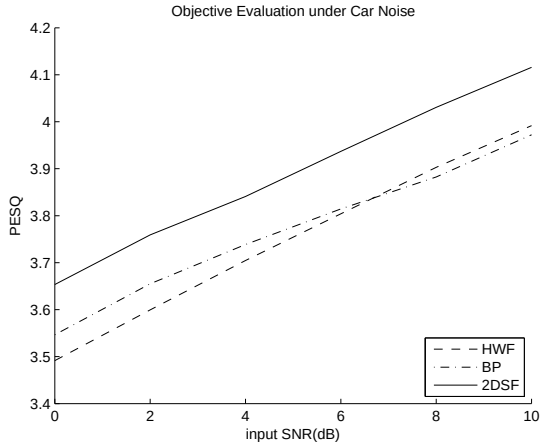

Figure 4: *PESQ measure under Car noise*

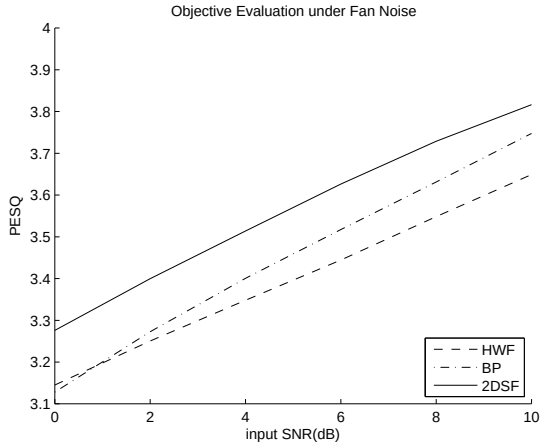

Figure 5: *PESQ measure under Fan noise*

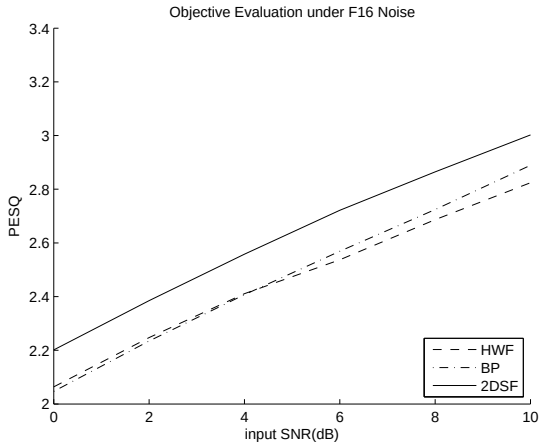

Figure 6: *PESQ measure under F16 aircraft noise*

hancement," in *Proceedings of HAVE*, 2003, pp. 61–64.

[4] I. Y. Soon and S. N. Koh, "Speech enhancement using 2-D Fourier transform," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 717–724, 2003.

[5] G. Whipple, "Low residual noise speech enhancement utilizing time-frequency filtering," in *Proc. ICASSP*, vol. 1, 1994, pp. I/5–I/8.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[7] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Improved noise suppression filter using self adaptive estimator of probability of speech absence," *Signal Processing*, vol. 75, pp. 151–159, 1999.

[8] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proceedings of INTERSPEECH-2006, Philadelphia, PA*, 2006.