Speech/Noise-Dominant Decision for Speech Enhancement

Sukhyun Yoon and Chang D. Yoo

Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea
Tel. : +82-42-869-5470, e-mail : ronald@eeinfo.kaist.ac.kr

Abstract

A novel method to reduce additive non-stationary noise is proposed. The proposed method requires neither the statistical assumption about noise nor the estimate of the noise statistics from any pause regions. The enhancement is performed on a band-by-band basis for each time frame. Based on both the decision on whether a particular band in a frame is speech or noise dominant and the masking property of the human auditory system, an appropriate amount of noise is reduced using modified spectral subtraction. The proposed method was tested on various noisy conditions - car noise, F16 noise, white Gaussian noise, pink noise, tank noise and babble noise. On the basis of comparing segmental SNR with spectral subtraction proposed by Boll with pause detection for estimating noise, and visually inspecting the enhanced spectrograms and listening to the enhanced speech, the proposed method was found to effectively reduce various noise while minimizing distortion to speech.

1. Introduction

In view of the steady rise in demand for various speech processing systems, the need for a high performance speech enhancement system ,when only a single channel of speech that has been degraded by additive noise is available, has increased. The addition of noise in speech reduces the recognition rate of a speech recognizer and decreases the coding efficiency of a vocoder. In general, noise reduces intelligibility and introduces listener fatigue. In the past, various methods for reducing noise have been proposed: spectral subtraction based method [1, 2], soft-decision filtering method [3], MMSE estimation method [4], model-based speech enhancement method [5], and enhancement method based on the human psychoacoustic masking property [6]. These methods require the statistical assumption of noise and when this assumption is not available, it is often estimated from pause regions that need to be detected with high accuracy - this may be difficult, and faulty detection increases the distortion of processed speech. Moreover, the noise estimate by pause detection assumes that the characteristic of noise changes slower than that of speech, therefore it cannot follow the variations of rapidly changing noise. The performance of an enhancement system hinges on the accuracy of the noise information. The proposed method requires neither the statistical assumption about noise nor the estimate of the noise statistics from any pause regions, and has a good performance in varying-noisy condition. There have been similar attempts to reduce noise without the pause detection[7, 8]; however, their results have been inconclusive or limited.

In this paper, the enhancement is performed on a band-by-band basis for each time frame. On the basis of both comparing the sum of spectral magnitude belonging to a particular band of a frame with that of previous frames and utilizing the masking[1] properties of the human auditory system, noise is reduced maximally while minimizing speech distortion. Specifically, the sums of previous frames are ordered in ascending order and are classified into two classes by the rate of increase in magnitude per frame. This ordered and classified sums determine whether the state of a particular band in a frame is speech dominant or noise dominant. Based on this decision and the masking property of the human auditory system, an appropriate amount[2] of noise is reduced using modified spectral subtraction.

The outline of this paper is as follows. Section 2 describes the overall enhancement system: Section 2.1 describes how to get a ordered sequence according to each critical band[3]. Section 2.2 describes how to classify each ordered sequence into two classes. Section 2.3 explains how to estimate noise spectrum and to enhance noisy speech. Section 3 presents some examples and experimental results using the proposed method. Finally, Section 4 concludes the paper.

2. Enhancement System

The proposed method enhances speech in three steps. First, noisy speech signal $y[n]$ is windowed by $w[n]$ and the windowed signal is transformed into DFT(Discrete Fourier Transform) coefficients. The coefficient magnitude pertaining to each critical band are summed, and the sums in each critical band for the previous L frames are sorted in ascending order. Second, using an approximation function, the ordered sequence is classified into two classes by the rate of increase in magnitude per frame. Each class has a different criterion for whether a particular band in a frame is speech or noise dominant. Third, based on both the decision on whether a particular band in a frame is speech or noise dominant and the masking property of the human auditory system, an appropriate amount of noise is reduced using modified spectral subtraction. Figure 1 shows the overall system (refer to section below for better understanding of figure).

[1] Masking is a process where one sound is rendered inaudible due to the presence of another sound.

[2] Because noise is masked by speech in speech dominant region, a small amount of noise estimate value is reduced for the preservation of speech.

[3] For a given frequency, the critical band is the smallest band of frequencies around it which activate the same part of the basilar membrane in human ear.

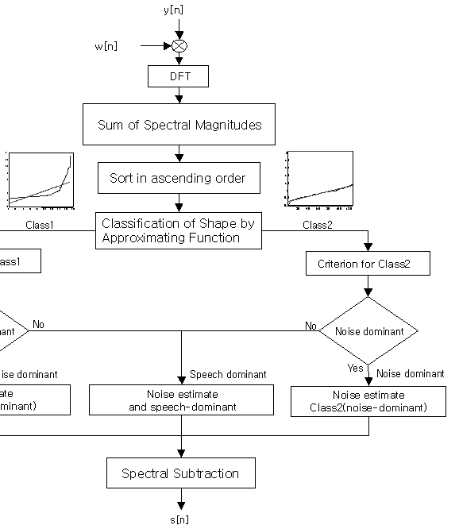

Figure 1: The Overall System.

2.1. Sorting Data

The spectral magnitude pertaining to the i-th critical band of the n-th frame are summed. This sum $A[i,n]$ is given by

$$A[i,n] = \sum_{k \in CB_i} |Y[k,n]| \qquad (1)$$

where CB_i is a set of frequency bins belonging to the i-th critical band and $Y[k,n]$ is the k-th DFT(Discrete Fourier Transform) coefficient for the n-th frame of noisy speech. For every i, length L sequence $\{A[i,j]\}_{j=n-L}^{n-1}$ is sorted in ascending order to obtain sequence $\{E[i,j]\}_{j=1}^{L}$, where $E[i,q]$ is the q-th largest term of $\{A[i,j]\}_{j=n-L}^{n-1}$. For example, $E[i,1] = \min_j(A[i,j])$ and $E[i,L] = \max_j(A[i,j])$ for $n-L \leq j \leq n-1$. The smaller the value of L, the faster the method adapts to varying noise. However, with smaller value L the reliability of the noise spectrum decreases. Figure 2 shows that the plot of ordered sequences $\{E[i,j]\}_{j=1}^{L}$ for i=5,10,15 where L=170.

2.2. The Classification of Ordered Sequence

The ordered sequence, $\{E[i,j]\}_{j=1}^{L}$, is classified into two classes by the rate of increase in magnitude per frame. When a considerable amount of speech is present in the last L frames, $\{E[i,j]\}_{j=1}^{L}$, the plot of $\{E[i,j]\}_{j=1}^{L}$ is a steep curve that shows a clear distinction between the values of noise and speech. *In this case,* $\{E[i,j]\}_{j=1}^{L}$ *is classified into Class 1.* When only noise and a small amount of speech are present in the last L frames, the plot of $\{E[i,j]\}_{j=1}^{L}$ is a flat curve that shows little distinction between the values of noise and speech. *In this case,* $\{E[i,j]\}_{j=1}^{L}$ *is classified into Class 2.*

A first order polynomial is used to fit the curve[9], and then the y-intercept of the polynomial is used to classify the ordered sequence $\{E[i,j]\}_{j=1}^{L}$. If the y-intercept is smaller than zero, $\{E[i,j]\}_{j=1}^{L}$ is classified into *Class 1*. If the y-intercept is larger than zero, $\{E[i,j]\}_{j=1}^{L}$ is classified into *Class 2*. Figure 3 shows the classification of the ordered sequences $\{E[i,j]\}_{j=1}^{L}$ for i=5 and 15 where L=170 into *Class 1* and *Class 2*.

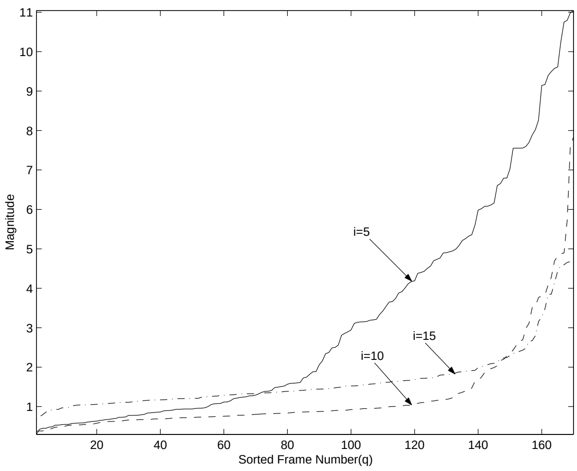

Figure 2: The plot of ordered sequences $\{E[i,j]\}_{j=1}^{L}$ for i=5,10,15 where L=170.

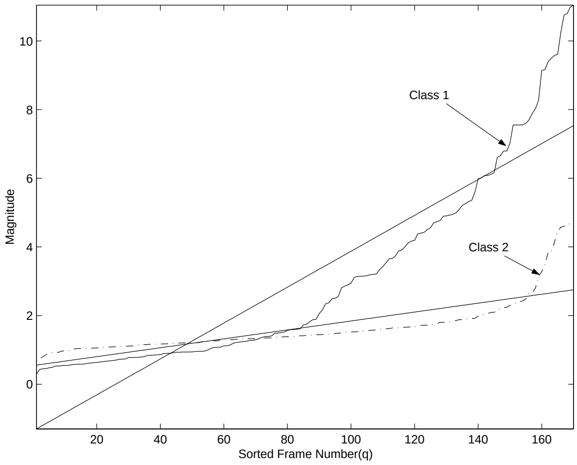

Figure 3: The classification of the ordered sequences $\{E[i,j]\}_{j=1}^{L}$ for i=5 and 15 where L=170 into either *Class 1* or *Class 2* by fitting a first order polynomial.

2.3. The Estimate of The Noise Spectrum

By comparing $A[i,n]$ with ordered $\{A[i,j]\}_{j=n-L}^{n-1}$ of last L frames, $\{E[i,j]\}_{j=1}^{L}$, the method determines whether (i,n)-*region*, the i-th critical band of the n-th frame, is speech-dominant or noise-dominant. If (i,n)-*region* is determined to be speech-dominant, the method uses a small component in $\{E[i,j]\}_{j=1}^{L}$ for $|N[i,n]|$ to subtract from $|Y[k,n]|$ where $k \in CB_i$, that is, to preserve speech. In this case, the residual noise is masked by the speech components. If (i,n)-*region* is determined to be noise-dominant, the method uses a large component in $\{E[i,j]\}_{j=1}^{L}$ for $|N[i,n]|$ to subtract from $|Y[k,n]|$ where $k \in CB_i$. For the enhancement of a particular band, the method considers the running statistics of previous L frames of that particular band and then determines whether a particular band in a frame is speech or noise dominant. The value for $|N[i,n]|$ is determined by the criterion set by each class. We can obtain enhanced speech by subtracting $|N[i,n]|$ from $|Y[k,n]|$. This is shown mathematically by

$$S[k,n] = rect(|Y[k,n]| - |N[i,n]|) \qquad k \in CB_i \qquad (2)$$

where $S[k,n]$ is the spectral magnitude of the enhanced speech at the n-th frame and $rect(\cdot)$ denotes half-wave rectification.

2.3.1. Criterion for Class 1

When $\{E[i,j]\}_{j=1}^{L}$ is classified as *Class 1*, it is assumed that there are strong speech components in a critical band for the last L frames. In terms of number, there are fewer strong speech components in the high-frequency band(e.g. $i > 17$) than in the low-frequency band(e.g. $i \leq 17$). Therefore, the distinction between speech and noise is clearer in the high-frequency band than in the low-frequency band, and the curve of $\{E[i,j]\}_{j=1}^{L}$ is steeper in the high-frequency band than in the low-frequency band. If $A[i,n]$ is smaller than the average of $E[i,q]$, that is, $\frac{1}{L}\sum_{q=1}^{L} E[i,q]$ in the high-frequency band, then (i,n)*-region* is noise-dominant.

In terms of number, there are more strong speech components in the low-frequency band(e.g. $i \leq 17$) than in the high-frequency band(e.g. $i > 17$). However, if (i,n)*-region* is a pause region, $A[i,n]$ is relatively small. Therefore, if $A[i,n]$ is smaller than threshold value, that is, $E[i, \lceil L \cdot a \rceil]$, where the range of a is from 0.25 to 0.35, [4] then, (i,n)*-region* is noise-dominant.

In the case stated above, $|N[i,n]|$ is estimated as a large value in $\{E[i,j]\}_{j=1}^{L}$. Otherwise, (i,n)*-region* is speech-dominant, and $|N[i,n]|$ is estimated as a small value in $\{E[i,j]\}_{j=1}^{L}$ for the preservation of speech, because noise is masked by speech in speech-dominant region. Noise spectrum is estimated as follows.

If ($A[i,n] < \frac{1}{L}\sum_{q=1}^{L} E[i,q]$ in high frequency)or
($A[i,n] < E[i, \lceil L \cdot a \rceil]$ in low frequency),
then
(i,n)*-region* is noise-dominant
$\Rightarrow \quad |N[i,n]| = E[i, \lceil L{\cdot}high \rceil]/B_i \;\; high \in [0.9, 1]$
otherwise, (i,n)*-region* is speech-dominant
$\Rightarrow \quad |N[i,n]| = E[i, \lceil L{\cdot}low \rceil]/B_i \quad low \in [0.25, 0.35]$

where B_i is the number of frequency bins in CB_i.

2.3.2. Criterion for Class 2

When $\{E[i,j]\}_{j=1}^{L}$ is classified as *Class 2*, it is assumed that there are only noise and weak speech components in the critical band for the last L frames. However, if $A[i,n]$ is especially large compared with the statistics for the last L frames in low frequency(e.g. $i \leq 17$), (i,n)*-region* is speech-dominant. Therefore,

If ($A[i,n]$ is in high frequency) or
($A[i,n] < E[i, \lceil L \cdot b \rceil]$ in low frequency), **then**
(i,n)*-region* is noise-dominant
$\Rightarrow \quad |N[i,n]| = c \cdot E[i, L]/B_i \qquad c \in (1, 2]$
otherwise, (i,n)*-region* is speech-dominant
$\Rightarrow \quad |N[i,n]| = E[i, \lceil L{\cdot}low \rceil]/B_i \;\; low \in [0.25, 0.35]$

where the range of b is from 0.9 to 0.99.

In Figure 4, (a) shows the plot of speech-dominant part in time-frequency domain as determined by the proposed method and Figure 4 (b) and (c) respectively show the spectrograms of noisy speech degraded by F16 noise (SNR=10dB) and enhanced speech.

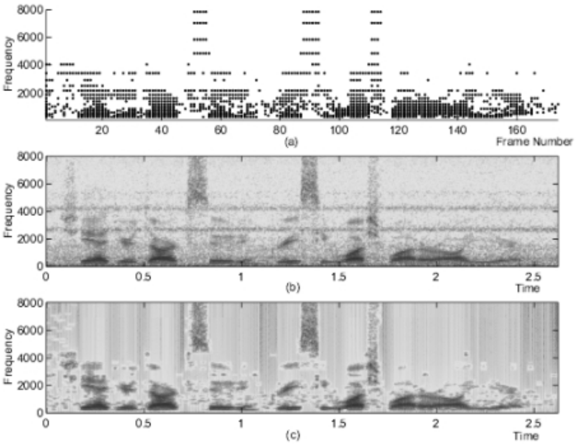

Figure 4: (a)The plot shows speech-dominant part in time-frequency domain as determined by the proposed method. The spectrograms of (b)noisy speech degraded by F16 noise(SNR=10dB) and (c)enhanced speech.

3. Evaluation

To illustrate the performance of the proposed method, we tested the method in various noisy conditions. Speech sentences from TIMIT database - ”She had your dark suit in greasy wash water all year” and ”Scholastic aptitude is judged by standardized tests” - respectively spoken by a male speaker and a female speaker, were used for evaluation. The following parameters have been chosen: 1)Hamming window of length N=512 (32ms) with 50% overlap; 2)total number of critical bands 22; 3)a=0.3, b=0.9, c=2, $high$=0.9, low=0.3; 4)L=50, 100.

Six different background noises were taken from the Noisex-92 database and were used for evaluation. The six noises were car noise, F16 noise, white Gaussian noise, pink noise, tank noise and babble noise.

Figure 5 shows the spectrograms of (a)noisy speech degraded by tank noise(SNR=10dB) (b)enhanced speech by spectral subtraction with pause detection and (c)enhanced speech by the proposed method(L=100). In Figure 5(b), result is very poor because the noise estimate by pause detection cannot follow the variations of tank noise. In Figure 5(c), noise is better reduced and speech is preserved well. The proposed method has a good performance for varying-noise.

Figure 6 (a) and (b) show respectively the spectrograms of noisy speech and enhanced speech(L=50). Noisy speech was obtained by degrading clean speech with F16 noise (SNR=10dB) for the duration 1.3 seconds followed by 1.3 seconds of car noise (SNR=-5dB) - ”She had your dark suit in greasy wash water all year” spoken by a male speaker. Although the statistical property of noise is changed abruptly, the proposed method adapts to varying-noise promptly.

Figure 7 shows the plots of segmental SNR improvement versus initial SNR for various noise conditions - car noise, white noise, F16 noise, tank noise, pink noise and babble noise.

In informal listening, the quality of the proposed system was superior to that of the spectral subtraction system with pause

[4] Notation ’$\lceil X \rceil$’ rounds the elements of X to the nearest integers towards infinity, e.g $\lceil 1.3 \rceil = 2$.

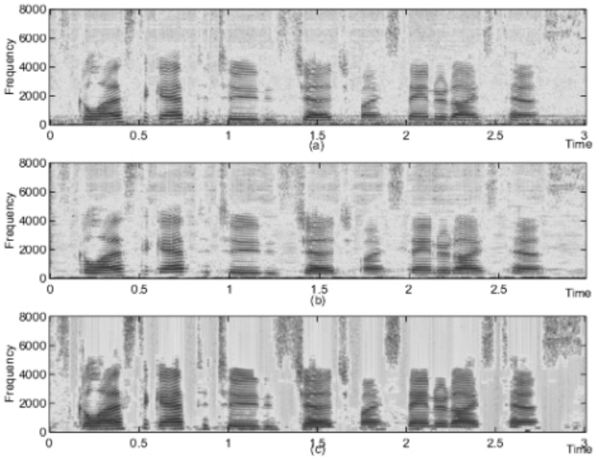

Figure 5: The spectrograms of (a)noisy speech degraded by tank noise(SNR=10dB) (b)enhanced speech by spectral subtraction with pause detection and (c)enhanced speech by the proposed method(L=100). ”Scholastic aptitude is judged by standardized tests” spoken by a female speaker

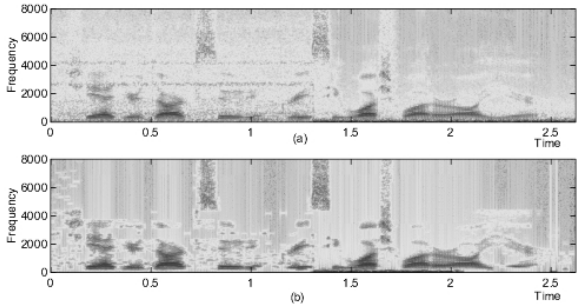

Figure 6: The spectrograms of (a)noisy speech and (b)enhanced speech. Noisy speech was obtained by degrading clean speech with F16 noise (SNR=10dB) for the duration 1.3 seconds followed by car noise (SNR=-5dB) for another 1.3 seconds. ”She had your dark suit in greasy wash water all year” spoken by a male speaker

detection. There were clearly fewer tonal noise in speech processed by the proposed system.

4. Conclusion

In this paper, we introduced a novel enhancement method that require neither the statistical assumption about noise nor the estimate of the noise statistics from any pause regions. The enhancement is performed on a band-by-band basis for each time frame. Based on both the decision on whether a particular band in a frame is speech or noise dominant and the masking property of the human auditory system, an appropriate amount of noise is reduced using modified spectral subtraction.

The proposed method was tested on various noisy conditions - car noise, F16 noise, white Gaussian noise, pink noise, tank noise and babble noise. On the basis of comparing segmental SNR with spectral subtraction proposed by Boll with pause detection for estimating noise, and visually inspecting the enhanced spectrograms and listening to the enhanced speech, the proposed method was found to effectively reduce various noise while minimizing distortion to speech.

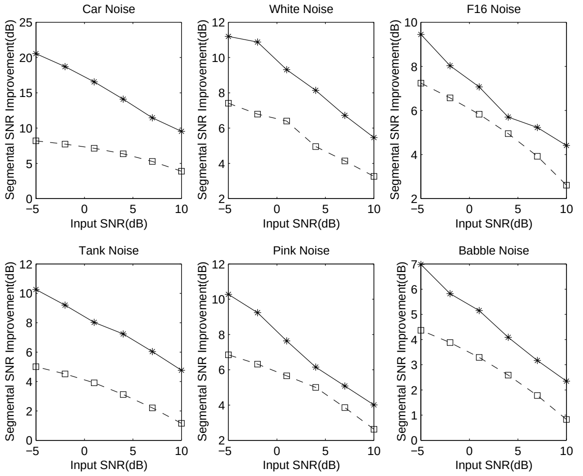

Figure 7: The plots of segmental SNR improvement versus initial SNR for various noise - car noise, white noise, F16 noise, tank noise, pink noise and babble noise. **solid line** : proposed method, **dashed line** : spectral subtraction with pause detection.

5. References

[1] S.F.Boll, “Suppression of acoustic noise speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 113–120, 1979.

[2] P.Lockwood and J.Boudy, “Experiments with a nonlinear spectral subtractor(nss), hidden markov models and projection, for robust recognition in cars,” *Speech Commun.*, pp. 215–228, June 1992.

[3] R.J.McAulay and M.L.Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 137–145, Apr. 1980.

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

[5] H.Sameti, H.Sheikhzadeh, and L.Deng, “Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE tran. on speech and audio processing*, pp. 445–455, Sep. 1998.

[6] N.Virag, “Single channel speech enhancement based on masking property of the human auditory system,” *IEEE tran. on speech and audio processing*, pp. 126–137, Mar. 1999.

[7] V.Stahl, A.Fischer, and R.Bippus, “Quantile based noise estimation for spectral subtraction and wiener filtering,” *ICASSP*, pp. 1875–1878, 2000.

[8] H.G.Hirsch and C.Ehrlicher, “Noise estimation techniques for robust speech recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 153–157, 1995.

[9] D.Kincaid and W.Cheney, *Numerical Analysis*, Brooks/Cole Publishing Company, 1996.