A NOVEL SPECTRAL SUBTRACTION SCHEME FOR ROBUST SPEECH RECOGNITON: SPECTRAL SUBTRACTION USING SPECTRAL HARMONICS OF SPEECH

Jounghoon Beh *Hanseok Ko*

Departments of Electronics and Computer Engineering, Korea University, Seoul, Korea
jhbeh@ispl.kore.ac.kr

ABSTRACT

This paper addresses a novel noise-compensation scheme to solve the mismatch problem between training condition and testing condition for the automatic speech recognition (ASR) system, specifically in the car environments. The conventional spectral subtraction schemes rely on the signal to noise ratio (SNR) such that attenuation is imposed on that part of the spectrum that appears to have low SNR, and accentuation is made on that part of high SNR. However, since these schemes are based on the postulation that the power spectrum of noise is in general at the lower level in magnitude than that of speech. Therefore, while such postulation is adequate for high SNR environment, it is grossly inadequate for low SNR scenarios such as car environment. This paper proposes an efficient spectral subtraction scheme focused to specifically low SNR noisy environments by distinguishing the speech-dominant segment from the noise-dominant segment in speech spectrum. Representative experiments confirm the superior performance of the proposed method over conventional methods. The experiments are conducted using car noise-corrupted utterances of Aurora2 corpus.

1. INTRODUCTION

The mismatch between training condition and testing condition is a major problem in ASR systems. The techniques to solve this problem can be categorized into two principal approaches. First is the spectral subtractive-type of algorithm performing noise suppression using short-time spectral amplitude, such as spectral subtraction, nonlinear spectral subtraction, and Weiner filter. The other is the feature compensation algorithm such as cepstral mean normalization or vector Tayler series. In general, it is well known that spectral subtractive-type algorithm is very simple and efficient especially in stationary noisy environments.

This paper is about a new spectral subtractive-type scheme based on the idea that even though speech is heavily corrupted by noise, the shape of spectral harmonics of speech is well preserved as when speech is not corrupted [6] [7].

The weakness of conventional spectral subtractive-type algorithm is identified and presented in Section 2. The proposed remedial approach is described in Section 3. In Section 4, we show the proposed method's effectiveness over conventional methods with representative experiments using Aurora 2. Concluding remarks are provided in Section 5.

2. SPECTRAL SUBTRACTIVE-TYPE ALGORITHM

When speech $x(n)$ is corrupted by background additive noise $b(n)$, the corrupted speech can be expressed as follows:

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

If speech and noise are assumed to be uncorrelated, in frequency domain, it can be represented as follows:

$$|Y(k)|^2 = |X(k)|^2 + |B(k)|^2 \tag{2}$$

where k is index of frequency bin.

2.1. Spectral Subtraction

In the case of power spectral subtraction, the short-time power spectrum $|\hat{X}(k)^2|$ of enhanced speech signal can be obtained by subtracting its noise estimate $|\hat{B}(k)^2|$ off from the corrupted speech. Note that every procedure is carried out in frame-by-frame basis [3]. Instead of power spectrum, its magnitude spectrum can be made available as follows.

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2 \left(1 - \alpha \dfrac{|\hat{B}(k)|^2}{|Y(k)|^2}\right), \\ \quad \mathit{if}\ |Y(k)|^2 - \alpha|\hat{B}(k)|^2 > \beta|Y(k)|^2 \\ \beta|Y(k)|^2, \\ \quad \mathit{otherwise} \end{cases} \tag{3}$$

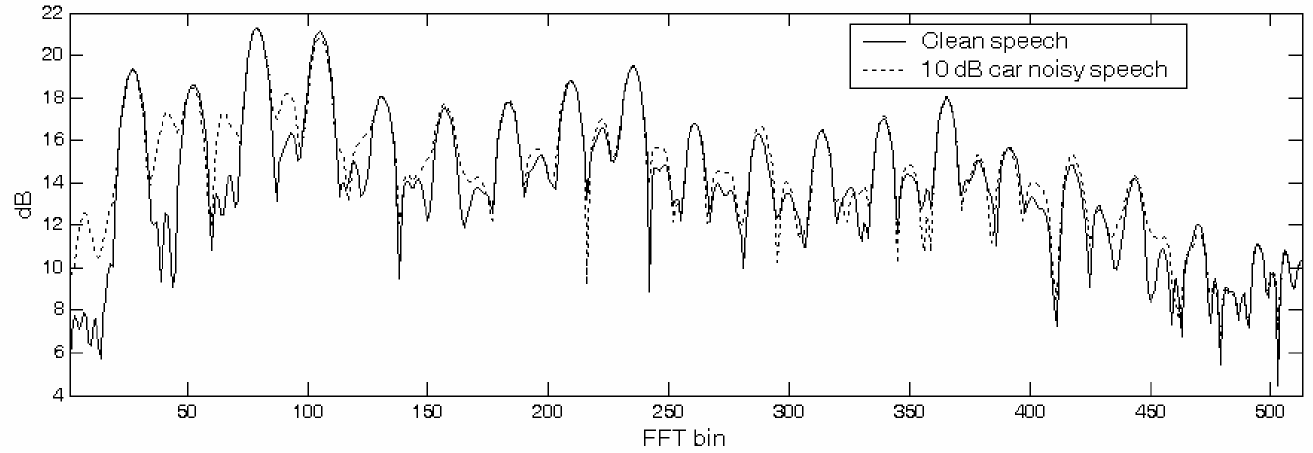

Figure. 1. Example spectrum of a speech frame (pronunciation /oh/ by female)

With over-subtraction factor α and floor factor β, this algorithm concerns the trade-off between noise reduction and residual noise. Note that the enhanced short-time power spectral amplitude $|\hat{X}(k)^2|$ depends on *a posteriori* SNR [5]:

$$\mathrm{SNR_p} = |Y(k)|^2 / |\hat{B}(k)|^2 \tag{4}$$

2.2. Nonlinear Spectral Subtraction

Nonlinear spectral subtraction algorithm is as follows [4]:

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2 \left(1 - \dfrac{|\hat{\Phi}(k)|^2}{|Y(k)|^2}\right), \\ \quad \mathit{if}\ \ |Y(k)|^2 - |\hat{\Phi}(k)|^2 > \beta |Y(k)|^2 \\ \beta |Y(k)|^2, \\ \quad \mathit{otherwise} \end{cases} \tag{5}$$

Function $\hat{\Phi}(k)$ can be chosen arbitrarily to implement the notion that relatively greater subtraction is applied to the low SNR region of spectrum and less subtraction to the high SNR region.

In general, the spectral amplitude of speech components (e.g. spectral harmonics) is higher than that of noise or the side lobes among harmonics so that those algorithms are rather suitable for reasonably high SNR (about 15~20dB) noisy speech cases. However, it appears that in the case of low SNR environments (especially 0~5dB), such that when the noise level is as close in amplitude as that of speech, there occurs unnecessarily subtracted speech region or less subtracted noisy region in noisy speech spectrum. Consequently, instead of the mundane use of SNR as a measure for subtraction, we need a new and better measure for activating the subtraction procedure. In particular, the subtraction over spectrum requires a more accurate measure than mere SNR in order to apply the subtraction rule, which is selective to speech-dominant region vs. noise-dominant region of spectrum.

3. PROPOSED ALGORITHM

3.1. Speech dominant region vs. noise dominant region

In speech, it is observed that the voiced speech segment has peaks positioned periodically in spectrum due to the vibration of vocal cords. Figure 1 illustrates with a sample spectrum of speech frame capturing the pronunciation /oh/ in one utterance contained in Aurora2 corpus. Note that at the peaks, their amplitudes are far greater than the amplitudes at the points between adjacent two peaks, or side lobes. Also, it is observed that the degree of corruption by the noise in the peaks is not as much as the degree of corruption at the points over the side lobes. From this observation, it can be deduced that in speech spectrum, the speech-dominant regions exist over or near the peaks and the noise-dominant regions exist over or near the side lobes.

The fundamental frequency can be obtained roughly by autocorrelation and then we can find the peak points from the roughly obtained fundamental frequency. For the frequency regions covering the peak points and their vicinity, we apply a small over-subtraction factor and large floor factor. In other regions, we apply a large over-subtraction factor and small floor factors. The detailed procedure is described in rest of Section 3. Note that all procedures in the proposed algorithm are done on a frame-by-frame basis.

3.2. Autocorrelation

Using autocorrelation methods, the fundamental frequency candidates are obtained [1]. Autocorrelation function is expressed as follows:

$$\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau) \tag{6}$$

3.3. FFT Analysis and Noise Estimation

Input speech is coded into 32ms frames, with a frame-shift of 16ms. Then short-time FFT is applied. In order to represent the spectrum precisely, we performed 1024 points FFT analysis on the relevant input frame.

For the noise estimation, we assume that any starting input speech is followed by a silence or background noise segment corresponding to 10 frames (176 ms). In noise duration, means of short-time spectral amplitude at each frequency bin are calculated.

3.4. Peak Points Detection

By means of log harmonic product spectra method [1], among the fundamental frequency candidates obtained by autocorrelation (section 3.1), one fundamental frequency f_0 is selected. In addition, the index of frequency bin k_0 corresponding to f_0 is calculated. Using k_0, the frequency axis is divided into several non-overlapping bands in the following form[6].

$$\left[\frac{2j-1}{2}k_0 \quad \frac{2j+1}{2}k_0\right], \quad j = 1,2,3... \tag{7}$$

Then, in each band, the harmonic components in each j-th band are assumed to be the frequency indices having the largest amplitude value in the relevant band. Through the above step, we finally determine the harmonic components in the spectrum of input frame, which are assumed to be speech-dominant region.

3.5. Voice Activity Detection (VAD)

This procedure is for the purpose of applying different subtraction rule to the speech frame and the non-speech frame. It is because of the non-speech frame, whose harmonics components cannot be determined. Consequently, the proposed scheme is inappropriate to non-speech frames.

Note that in each input frame, whether it is the speech frame or not, the fundamental frequency f_0 is calculated. As a result, the value of autocorrelation about selected f_0 is also calculated at each frame inevitably. Two separate procedures are employed for robust detection of speech vs. non-speech region. First, we take a logarithm to this value, and then a smoothing procedure is carried out. If this value is greater than half of the mean value of autocorrelation obtained during the first 10 frames, the relevant frame is decided as ‘speech frame’. Secondly, for frames determined as ‘non-speech’, if the value of autocorrelation is over 0.0015 * $\phi(0)$ and also the detected f_0 value is in 50~400Hz, then it is concluded as ‘speech frame’ once again. It is well known that 50~400Hz is considered an appropriate fundamental frequency range for human voice.

3.6. Spectral Subtraction

Based on the result of VAD procedure, we apply different subtraction rule.

3.6.1 Speech frame

In order to implement the proposed scheme, we designed following simple linear function.

- If $k \in \left[\frac{2j-1}{2}k_0 \quad k_j\right], \quad j = 1,2,3...$, then

$$\gamma(k) = \frac{\alpha_{MAX} - \alpha_{\min}}{k_j - \frac{2j-1}{2}k_0}(k_j - k) + \alpha_{\min} \tag{8}$$

$$\delta(k) = \frac{\beta_{\min} - \beta_{MAX}}{k_j - \frac{2j-1}{2}k_0}(k_j - k) + \beta_{MAX} \tag{9}$$

- If $k \in \left[k_j \quad \frac{2j+1}{2}k_0\right], \quad j = 1,2,3...$, then

$$\gamma(k) = \frac{\alpha_{MAX} - \alpha_{\min}}{\frac{2j+1}{2}k_0 - k_j}(k - k_j) + \alpha_{\min} \tag{10}$$

$$\delta(k) = \frac{\beta_{\min} - \beta_{MAX}}{\frac{2j+1}{2}k_0 - k_j}(k - k_j) + \beta_{MAX} \tag{11}$$

$\gamma(k)$ applies the minimum over-subtraction factor to each harmonic component and the maximum over-subtraction factor to the middle point at each components. Then over-subtraction factors of points exist in the interval between those points are interpolated linearly. Figure.2. illustrate a shape of $\gamma(k)$ which of fundamental frequency is about 156Hz when sampling rate is 8000Hz and 1024 points FFT analysis is applied.

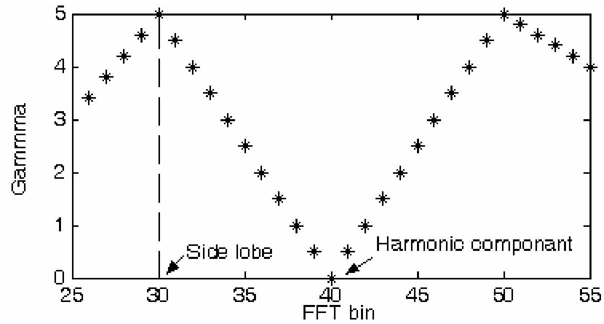

Figure. 2. An illustration about implementing over-subtraction factor to FFT bins

Also, reverse process is implemented with respect to frequency axis and harmonic components using $\delta(k)$.

The values of parameters used are as follows:

$$\alpha_{MAX} = 5, \alpha_{\min} = 0, \beta_{MAX} = 0.2, \beta_{\min} = 0.05 \tag{12}$$

3.6.2 Non-speech frame

We applied the subtraction rule as same as conventional method using Eq. (1) with $\alpha_{MAX}, \beta_{\min}$ in Eq. (12).

However, in speech frame, if there is the band of which SNR value is below 7dB, in that band, the subtraction rule for the non-speech frame is applied. SNR is defined as follows:

$$SNR = 10\log\left(\frac{Y^2(k_j) - \hat{B}^2(k_j)}{\hat{B}^2(k_j)}\right) \tag{13}$$

4. EXPERIMETNAL RESULTS

4.1. Experimental Conditions

In these experiments, utterances corrupted by car noise among Aurora2 corpus are used. For comparison, spectral subtraction algorithm [2] and nonlinear spectral subtraction algorithm [3] are evaluated. Throughout all experiments, we used 1024 points FFT analysis as same as that of the proposed method. Finally, for each algorithm, the enhanced speech signals are recovered by overlap-and-add manner followed by inverse FFT on frame-by-frame basis. All performances are evaluated.

4.2. Experimental Results

Type SNR	Baseline	SS	NSS	Proposed
clean	99.02	98.90	98.87	98.81
20dB	97.97	98.57	98.63	98.06
15dB	94.24	97.55	97.58	96.99
10dB	78.17	93.47	93.59	93.20
5dB	42.59	75.22	75.72	82.11
0dB	14.67	39.16	39.25	52.67
-5dB	9.25	12.41	12.05	18.19
Avg.	69.79	80.79	80.95	84.60

Table 1. Word Accuracy (%)

In Table 1, 'SS' denotes the spectral subtraction method and 'NSS' to the nonlinear spectral subtraction and 'Avg' to the mean value of word accuracy over 0dB~20dB. From the Table, the proposed method is seen effective at low SNR cases. The average word accuracy of the proposed method also shows that it is superior over other spectral subtraction approaches.

5. CONCLUSIONS

This paper proposes an efficient spectral subtraction scheme focused to specifically low SNR noisy environments by distinguishing the speech-dominant segment from the noise-dominant segment in speech spectrum. Representative experiments confirm the superior performance of the proposed method over conventional methods. In particular, the proposed method is seen effective at low SNR cases (SNR < 5 dB). The average word accuracy of the proposed method also shows that it is superior over other recently introduced approaches. The experiments are conducted using car noise-corrupted utterances of Aurora2 corpus.

6. REFERENCES

[1] L.Rabiner, R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall 1978.

[2] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transaction on Acoustics, Speech and Signal Processing* , Vol.27, No.2, 113-120, April 1979

[3] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by additive noise," *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 208-211, April 1979.

[4] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor(NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication*, Vol. 11, pp. 215-28, 1992

[5] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 2, 126-137, Mar 1999.

[6] J. Jensen, J. Hansen, "Speech Enhancement Using a Constrained Iterative Sinusoidal Model," IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 7, Oct 2001.

[7] D. Ealey, H. Kellher, D. Pearce, "Harmonic tunneling: tracking non-stationary noises during speech," *Eurospeech 2001*, 437-440, 2001