A NOVEL MULTI-BAND SPECTRAL SUBTRACTION METHOD BASED ON PHASE MODIFICATION AND MAGNITUDE COMPENSATION

Chao Li, and Wen-Ju Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
E-mail: {cli, lwj}@nlpr.ia.ac.cn

ABSTRACT

The spectral subtraction method is historically one of the first algorithms proposed for noise reduction. However, most implementations and variations of the basic technique cannot suppress musical noise effectively, and keep residual noise and speech distortion under a low level simultaneously. In this paper, we analyze two foundational shortcomings of spectral-subtraction-like methods, and proposed a novel multi-band spectral subtraction method based on phase modification and magnitude compensation. The new method outperforms the standard multi-band spectral subtraction method resulting in superior speech quality and greatly reduced musical noise.

Index Terms— Multi-band spectral subtraction, phase modification, magnitude compensation, musical noise

1. INTRODUCTION

The spectral subtraction (SS) method, proposed by Boll [1], has been extensively studied because of its simplicity and effectiveness. Its basic principle is to subtract the spectrum of noise from that of the noisy speech. Although the spectral subtraction algorithm can be easily implemented to effectively reduce the noise present in the corrupted signal, it introduces a perceptually annoying artifact which is commonly referred to as musical noise.

Oversubtractive rule is widely used in improved spectral subtraction method [2-5] to reduce musical noise by subtracting an overestimate of noise spectrum, while preventing the resultant spectral components from going below a present minimum value. The level of perceived musical noise is reduced, but background noise remains and target speech is distorted. Therefore, it is desirable to develop an algorithm that reduces musical noise without adverse impact. The fundamental factors that degrade the performance of spectral subtraction method are:

1) Without phase information of noise. Although the signal phase need not to be recovered exactly as long as the spectral signal-to-noise (SNR) is high enough, phase estimation is critically important for accurate magnitude estimation, especially under low-SNR condition.

2) Inaccurate estimate of the noise spectrum. The significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum results in the presence of isolated residual noise levels of large variance.

It is hard, even impossible, to solve the above-mentioned factors directly. In this paper, we propose a novel spectral subtraction method to overcome them indirectly. Firstly, phase modification algorithm eliminates the adverse effect of noise phase, and obtains the maximum magnitude of the noisy speech with the phase of target speech equal to that of noise. Furthermore, suitable subtractive rule is chosen in each frequency band to subtract just the necessary amount of noise spectrum. Finally, magnitude compensation algorithm enhances the attenuated speech caused by oversubtraction and rough estimate of the noise spectrum.

2. IMPROVED SPECTRAL SUBTRACTION

Assume that $y(n)$, the noise-corrupted input signal, is composed of the clean speech $x(n)$ and the additive noise $v(n)$,

$$ y(n) = x(n) + v(n). $$  \hspace{1cm} (1)

Taking the $N$-point discrete Fourier transform (DFT) gives

$$ Y(k) = X(k) + V(k). $$  \hspace{1cm} (2)

In polar form,

$$ \alpha_y e^{j\theta_y} = \alpha_x e^{j\theta_x} + \alpha_v e^{j\theta_v}. $$  \hspace{1cm} (3)

where $\alpha_y$ is the magnitude spectrum, and $\theta_y$ is the phase spectrum of the corrupted signal. $\alpha_x$, $\alpha_v$, $\theta_x$ and $\theta_v$ are defined in a similar way. We drop the frequency bin $k$ for convenience. $Y$ can be represented geometrically in the complex plane as the sum of two complex numbers, $X$ and $V$, as illustrated in Fig.1. The noisy signal magnitude $\alpha_y$ is obtained by the
following equation:

$$\alpha_y^2 = \alpha_x^2 + \alpha_v^2 + 2\alpha_x \alpha_v \cos \theta_{vx}.$$  

where $\theta_{vx} = \theta_v - \theta_x$. Considering the generalized version of the spectral subtraction algorithm,

$$\hat{\alpha}_x^p = \alpha_y^p - \alpha_v^p,$$  

where $p$ is the power exponent, with $p = 1$ yielding the magnitude spectral subtraction, and $p = 2$ yielding the power spectral subtraction. Since $\alpha_v$ cannot be directly obtained, an estimate $\hat{\alpha}_v$ is calculated during periods of silence.

Although the clean and noise signal are statistically uncorrelated, the cross terms in (4) might not necessarily be 0 or 1 in each signal segment (20-30 ms). Therefore, the presence of cross terms brings variance to spectral subtraction.

### 2.1. Phase modification algorithm

How to eliminate the adverse effect of cross terms? If we find a time point around the processing speech segment whose difference phase $\theta_{vx} = 0$, the standard magnitude subtraction rule should be very accurate.

Assume the phase difference $\theta_{vx}$ is a random variable uniformly distributed in $[-\pi, \pi]$, and the speech phase continuous change. Defining event:

$$A : \theta_{vx} = 0 \text{ in current time point.}$$  

Then, its probability is

$$P(A) = P(\theta_{vx} = 0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f = \frac{f}{f_s},$$  

where $2\pi f/f_s$ is the difference phase between adjacent time points, $f$ denotes the frequency, $f_s$ is sample rate. Let $A_1$ be the event that $A$ occurs at least once in $M$ continuous time points. Its probability is

$$P(A_1) = 1 - P(A^c)^M = 1 - (1 - f/f_s)^M = 1 - (1 - k/N)^M,$$  

where $A^c$ is the complement of $A$, $k$ is the frequency bin, $N$ is the size of DFT. We hope such probability no less than 0.9, then the number of time point $M$ can be obtained by:

$$M(k) = \left\lfloor \log_{1 - k/N} 0.1 \right\rfloor_{\text{even}},$$  

where $\lfloor \cdot \rfloor_{\text{even}}$ operator gets the nearest even less than operand. Taking $f_s = 8$ KHz ($N = 512$) and $f_{\text{min}} = 60$ Hz ($k_{\text{min}} = 4$) as an example, $M_{\text{max}} = 306$.

Therefore, while processing a signal segment, we search from $M(k)/2$ past to $M(k)/2$ future time points for each frequency bin $k$ to obtain the noisy speech magnitude estimate $\hat{\alpha}_y$ with $\theta_{vx} = 0$:

$$\hat{\alpha}_y(k) = \frac{\max_{M(k)/2 \leq i \leq M(k)/2} |Y_i(k)|}{\abs{Y_i(k)}}.$$  

where $Y_i$ is the DFT of $y(n-i)$, which contains the $L$ most recent samples of the noisy speech signal at time point $(n-i)$:

$$y(n-i) = [y(n-i), \cdots, y(n-i-L+1)]^T.$$  

### 2.2. Multi-band magnitude spectral subtraction

Original multi-band spectral subtraction (MBSS) method was proposed in [3] for power spectrum. It is based on the fact that noise affects the speech spectrum differently at various frequencies. The speech spectrum is divided into $R$ nonoverlapping bands, and spectral subtraction is performed independently in each band [3]. We can estimate the clean speech magnitude spectrum in the $i$th band by:

$$\hat{\alpha}_{x,i}(k) = \hat{\alpha}_{y,i}(k) - \phi_i \cdot \delta_i \cdot \hat{\alpha}_{v,i}(k), \quad b_i \leq k \leq e_i,$$  

where $b_i$ and $e_i$ are the beginning and ending frequency bins of the $i$th band, $\phi_i$ is the oversubtraction factor of the $i$th band, and $\delta_i$ is an additional band-subtraction factor, which can provided an additional degree of control within each band. The values for $\delta_i$ are empirically determined and set to:

$$\delta_i = \begin{cases} 
1 & e_i \leq N \cdot 1000/f_s \\
1.5 & N \cdot 1000/f_s < e_i \leq N/2 - N \cdot 3000/f_s \\
2 & e_i > N/2 - N \cdot 3000/f_s 
\end{cases}$$  

(13)

Negative values resulting from the subtraction in (12) are floored to the noisy spectrum as follow:

$$\hat{\alpha}_{x,i} = \begin{cases} 
\hat{\alpha}_{x,i} & \text{if } \hat{\alpha}_{x,i} > \phi \cdot \hat{\alpha}_{y,i} \\
\phi \cdot \hat{\alpha}_{y,i} & \text{else}
\end{cases}$$  

(14)

where the maximum attenuation threshold $\phi$ is set to 0.02. The band-specific oversubtraction factor $\phi_i$ is a function of the segmental $SNR_i$ of the $i$th band which is calculated by:

$$\varphi_i = \begin{cases} 
3 & SNR_i \leq -5 \\
2.6 - 0.08 \times SNR_i & -5 < SNR_i \leq 20 \\
1 & SNR_i > 20
\end{cases}$$  

(15)

where the band $SNR_i$ is given by:

$$SNR_i = 10 \log_{10} \left( \frac{\sum_{k=b_i}^{e_i} |\hat{\alpha}_{y,i}(k)|^2}{\sum_{k=b_i}^{e_i} |\hat{\alpha}_{v,i}(k)|^2} \right)$$  

(16)
2.3. Magnitude compensation algorithm

In order to further mask any remaining musical noise and enhance the attenuated speech component, a suitable amount of noise spectrum is introduced back to enhanced spectrum. This algorithm can be realized as follows:

$$\tilde{\alpha}_{x,i} = \tilde{\alpha}_{x,i} + \mu_{1,i} \cdot \tilde{\alpha}_{y,i} + \mu_{2,i} \cdot \tilde{\alpha}_{y,i}^2$$ (17)

where the first-order compensation factor $\mu_{1,i}$, set to 0.05, is used to mask residual musical noise; the second-order compensation factor $\mu_{2,i}$ contributes a stronger compensation than $\mu_{1,i}$, whose value is correlated with the segmental $SNR_i$ of the $i$th band. In low-SNR condition ($\leq 0$dB) the speech maybe absent, so no compensation is needed; when SNR is high enough (> 10dB), in contrast, speech magnitude is so enough that compensation is not needed either. Therefore, with enhanced spectrum with middle-SNR needs compensation urgently. The value of $\mu_{2,i}$ is a function of the segmental $SNR_i$, which can be empirically determined as follows:

$$\mu_{2,i} = \begin{cases} 0 & \text{if } SNR_i < SNR_0 \frac{SNR_i - SNR_0}{4} \exp \left( - \frac{(SNR_i - SNR_0)^2}{8} \right) & \text{else} \end{cases},$$ (18)

where $SNR_0$ is threshold of $SNR_i$ to distinguish voiced activity from silence, which is set to 3dB.

The effect of both compensation factors, $\mu_{1,i}$ and $\mu_{2,i}$ is illustrated in Fig.2 (segmental SNR≈ 5dB, white noise). $\mu_{1,i}$ and $\mu_{2,i}$ (especially $\mu_{2,i}$) benefit the enhanced spectrum with a similar strong peak in middle-SNR frequency bands, while a low valley in low-SNR frequency bands. By comparing to original MBSS method, our proposed method can achieve less speech distortion and residual noise.

The NOIZEUS [7] speech corpus was used to evaluate the proposed method. The noisy database contains 30 sentences produced by three male and three female speakers, and was corrupted by eight different real-world noises at different SNRs. The sample rate is 8 kHz. As mentioned in [3], it is the optimal choice for speech quality and perception in the case of 8 linearly spaced bands.

For evaluation, the SNR and log-likelihood ratio (LLR) (marked in $[0, 2]$) [8] are adopted as the objective measures to denote noise reduction and speech distortion respectively. Simultaneously, perceptual evaluation of speech quality (PESQ) (marked in $[0.5, 4.5]$) [9] measure is used to replace the subjective listening test due to its high correlations ($\rho > 0.92$) [9]. High speech quality is denoted by high values of the SNR and PESQ, and low value of the LLR.

Fig.4 shows the comparative results between the proposed method and the original MBSS method in database at SNR = 5dB. As can be seen, compared with original MBSS in all types of noises, the proposed method improves SNR measure by an average of 1.7dB, while remains the similar average value of the LLR measure with a fluctuation less

![Fig. 2. The effect of compensation factors with the different enhanced signals.](image)

3. IMPLEMENTATION

The black diagram of the proposed multi-band spectral subtraction based on phase modification and magnitude compensation is shown in Fig.4. The signal is first hamming windowed using a 30-ms window and 50% overlap. The magnitude spectrum is estimated using the Fast Fourier Transform (FFT). The noisy speech spectrum is then preprocessed to produce a maximum estimate by using phase modification algorithm. Next, the noise and speech spectra are divided into $R$ adjacent bands and the segmental SNR is calculated in each band. Oversubtraction factor and magnitude compensation second-order factor are computed for each band. The estimated noise in each band is subtracted form the noisy speech spectrum. Finally, the processed bands are recombined and the enhanced signal is obtained by IFFT using the noisy speech phase.

![Fig. 3. Block diagram of proposed algorithm.](image)
than 5%. PESQ measure improves an average of 0.13, which confirms that the proposed method resulted in better quality and perceptual evaluation.

Compared between the various noises, the proposed method achieves a better performance in some types of noise, such as car, exhibition, train, station and street, while a relative mediocre performance in other types of noise. It is the speech-shape noise that causes such performance degradation, due to its almost continuously variable phase and rapidly changing magnitude. These characters of speech-shape noise decay phase modification algorithm and magnitude compensation algorithm to some extent.

To compare between the various input SNRs, we take the database corrupted by car noise as an example (as showed in Tab.1). Most of the values indicate again that the proposed method has an improved speech quality.

### 5. CONCLUSION

In this paper, a novel spectral subtraction method was presented for enhancing speech corrupted by real-world noises, which based on three essential algorithms: phase modification algorithm, multi-band oversubtraction rule and magnitude compensation algorithm. These algorithms make up for foundational deficiencies in the spectral subtraction method. Results showed that the proposed method outperformed the conventional multi-band spectral subtraction.

### 6. REFERENCES


