A Speech Enhancement Method by Coupling Speech Detection and Spectral Amplitude Estimation

Feng Deng, Chang-chun Bao, Feng Bao
Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
dengfeng@emails.bjut.edu.cn, baochch@bjut.edu.cn, baofeng@emails.bjut.edu.cn

Abstract

In this paper, a single-channel speech enhancement method by coupling speech detection and spectral amplitude estimation is proposed. First, the optimal detector used for the spectral coefficients of the speech signal is obtained by minimizing the combined risk function which considers both detection and estimation errors. Second, according to the optimal speech detector, the optimal spectral amplitude estimator is obtained by further minimizing the combined risk function. For speech-presence and speech-absence, we propose two general weighted cost functions which are based on the human auditory perception and compressive nonlinearities of the cochlea. The cost parameters are employed to balance the speech distortion and residual noise caused by missed detection and false alarm, respectively. The performance of the proposed method is evaluated under ITU-T G.160. The experimental results indicate that the proposed method evidently performs better than the reference methods.

Index Terms: speech enhancement, speech detection and estimation, combined risk function

1. Introduction

Speech enhancement has been studied for several decades, and a large number of speech enhancement approaches are already proposed, such as spectral-subtraction method [1], the MMSE method of Ephraim-Malah [2], log spectral amplitude (LSA) estimator [3], speech enhancement based on speech-presence uncertainty [4]. These methods often operate in the discrete Fourier transform (DFT) domain, in which the speech spectral coefficients are estimated from the noisy speech.

As we know, speech is present only in some signal frames, and only some frequency bins contain the significant speech energy in each frame. Namely, the spectral coefficients of the speech signal are generally sparse. However, the existing speech enhancement methods do not take the sparse characteristics into consideration, which often focus on estimating the spectral coefficients rather than detecting the existence of speech signal. The spectral-subtraction method [1] can detect the existence of speech by signal power in the frequency domain, but it is so simple that it produces music on the speech-presence uncertainty can yield reasonable estimators [2] based on speech-presence uncertainty, which may degrade since the change of the noise power which makes the speech probability calculated falsely.

The significance of signal detection and estimation for speech enhancement, a simultaneous detection and estimation approach (SDEA) for speech enhancement is presented in [5], which includes the operations of detection and estimation simultaneously. However, the spectral amplitude (QSA) error is used as its cost function which limits the ability of noise reduction and affects the performance of the method. Therefore, jointly considering the important influence of the signal detection and cost function for speech enhancement, in this paper, we propose a single-channel speech enhancement method by coupling speech detection and spectral amplitude estimation, in which the operations of speech detection and spectral amplitude estimation are strongly coupled. The optimal detector used for the spectral coefficients of the speech signal and the optimal spectral amplitude estimator are achieved by jointly minimizing the combined Bayes risk function which considers both detection and estimation errors. For speech-presence and speech-absence, in order to take the advantage of both perceptual weighting and cochlea’s compressive nonlinearities, we present two general weighted cost functions which jointly use the weighted factor $p$ in [6] and power exponent $\beta$ in [7]. Furthermore, the cost parameters are used to balance the speech distortion and residual noise caused by missed detection and false alarm, respectively. Experimental results show that the proposed method yields better performance than reference methods.

The remainder of this paper is organized as follows. In Section 2, the speech enhancement method by coupling speech detection and spectral amplitude estimation is described. In Section 3, we present the amplitude estimators based on the general weighted cost functions. The performance evaluation is presented in Section 4, and Section 5 gives the conclusions.

2. Coupling speech detection and spectral amplitude estimation

It is assumed that the clean speech signal $x(n)$ is contaminated by an uncorrelated additive noise $d(n)$, then the noisy speech signal $y(n)$ can be expressed as: $y(n) = x(n) + d(n)$. Applying the Fourier transform to both sides of $y(n)$, we can get:

$$Y(\omega_k) = X(\omega_k) + D(\omega_k)$$  \hspace{1cm} (1)

where $Y(\omega_k)$, $X(\omega_k)$ and $D(\omega_k)$ are the Fourier transform coefficients of noisy speech, clean speech and noise signal, respectively. $\omega_k = 2\pi k/N$, $k \in [0, N)$ is the index of frequency bin, $N$ is the frame length.

Since the human auditory system is not sensitive to the phase spectrum, we can replace the phases of clean speech and noise signal by the one of the noisy speech, then we can rewrite (1) in polar form:

$$Y e^{j\phi_k} = X e^{j\phi_k} + D e^{j\phi_k} = (X + D) e^{j\phi_k}$$  \hspace{1cm} (2)
where \( Y_k, X_k \) and \( D_k \) denote the \( k \)-th magnitudes of the noisy speech, clean speech and noise signal, respectively, and \( \theta(j,k) \) is the phase spectrum of noisy speech.

According to the existence of speech signal in DFT coefficients, we define two state spaces \( H^i \) (\( i=0,1 \):
\[
H^0 : Y_k = D_k \\
H^1 : Y_k = X_k + D_k
\]
where \( H^0 \) and \( H^1 \) denote the speech is absent and present in the \( k \)-th DFT coefficients, respectively.

In addition, two decision spaces \( \eta^j \) (\( j=0,1 \)) are defined for detecting the existence of speech signal. Under the decision \( \eta^j \), the state space \( H^j \) is detected and the corresponding enhanced speech \( \hat{X} = \hat{X}_{k,0} \) is obtained. Under the decision \( \eta^0 \), the speech state \( H^0 \) is accepted and the corresponding speech estimation \( \hat{X} = \hat{X}_{k,1} \) is achieved. We assume the cost function \( C_{k}(X_k, \hat{X}) \) denote the cost of making a decision \( \eta^j \), and then the combined Bayes risk function \( R \) which couples speech detection and spectral amplitude estimation is shown in Fig. 1.

Figure 1: The block diagram of speech enhancement method by coupling detection and estimation

Therefore, the optimal spectral amplitude estimator \( \hat{X}_{k,j} \) under the optimal speech decision \( \eta^j \) is obtained from (8):
\[
\hat{X}_{k,j} = \arg \min_d \{ q_r(Y(\omega)) + (1-q)q_c(Y(\omega)) \}
\]
(10)

3. The spectral amplitude estimators based on the general weighted cost function

3.1. Spectral amplitude estimator

From (9) and (10) we can see that, both the optimal speech detector and spectral amplitude estimator contain the risk \( r_c(Y(\omega)) \) which depends on the cost function \( C_{k}(X_k, \hat{X}) \). And the cost function plays a significant role in the Bayesian estimators. For different cost functions, we can get various kinds of amplitude estimators. Therefore, we define the cost function associated with the state-decision pair \( \{H^0, \eta^j \} \):
\[
C_{k}(X_k, \hat{X}) = c_{j}d_{k}(X_k, \hat{X})
\]
(11)
where \( c_j \) is the cost parameter which balances the costs associated with the state-decision pair \( \{H^0, \eta^j \} \). \( c_{00} = c_{11} = 1 \) denotes the decision is correct; \( c_{0j} = 5 \) denotes the cost of false alarm (i.e. the speech absent is detected as speech presence); \( c_{10} = 5 \) represents the cost of missed detection (i.e. the speech presence is detected as speech absent). For the speech state space \( H^j \) (\( i=0,1 \)), the general weighted distortion measure \( d_{k}(X_k, \hat{X}) \) is defined as follows:
\[
d_{k}(X_k, \hat{X}) = \int_{\Omega} \int_{\Omega} \{ G_{k}(X_k - \hat{X})^{\beta} \} R_d(Y(\omega)) + (1-q)R_c(Y(\omega)) \}
\]
(12)
where \( G_{k} \) denotes gain floor, \( p \) is the weighted factor, \( \beta \) is the power exponent. Under speech state \( H^j \), the distortion measure takes the advantage of both perceptual weighting [8] and cochlea’s compressive nonlinearities [3]. Under speech state \( H^0 \), the distortion measure is defined to allow some comfort background noise level in enhanced speech signal.

Assuming both \( X(\omega) \) and \( D(\omega) \) are zero-mean, complex Gaussian variables with variances \( \lambda_{X} \) and \( \lambda_{D} \) respectively. By substituting (11) and (12) into (10), we can get
\[
\hat{X}_{k,j} = \arg \min_d \{ q_r(Y(\omega)) + (1-q)q_c(Y(\omega)) \}
\]
(13)
where \( q_r(Y(\omega)) \) depends on the spectral amplitude estimation \( \hat{X}_{k,j} \) through the cost function \( C_{k}(X_k, \hat{X}) \). The block diagram of the speech enhancement method by coupling detection and estimation is shown in Fig. 1.

Taking the derivation of (13) with respect to \( \hat{X} \) and setting it equal to zero, then using Bayes’ criterion we can get:
\[
\hat{X}_{k,j} = \frac{c_{j}A(Y)}{c_{j}A(Y)} \int_{\Omega} X_k \hat{X}^{\beta} p(X_k | Y(\omega)) d\Omega X_k + c_{0j}(G_{k} \hat{X})^{\beta}
\]
(14)
where \( A(Y) = \frac{q - p(Y(\eta_j)|H_j)}{1 - q - p(Y(\eta_j)|H_j)} \) is the generalized likelihood ratio.

For (14), using the Gaussian statistical model [6], we can derive the optimal estimator \( \hat{X}_j \) under the optimal decision \( \eta'_j \):
\[
\hat{X}_j = \frac{c_1 A(Y)}{\lambda_1} \left[ \frac{\sqrt{\lambda_1}}{\lambda_1} \frac{\Gamma(\frac{p+\beta}{2} - 1)}{\phi(\frac{p+\beta}{2} - 1)} \right]^{(\frac{p+\beta}{2} - 1)} \] (15)
where \( G_{ij} \) is the function of the optimal decision \( \eta'_j \), \( \Gamma(\cdot) \) is the Gamma function, \( \Phi(\cdot) \) denotes the confluent hypergeometric function, \( \zeta_0 \) is a priori SNR, \( \eta_0 \) is a posteriori SNR, \( \nu_0 \) is the function of \( \zeta_0 \) and \( \eta_0 \), which are defined in [6].

In addition, from (9) we can find that, in order to obtain the optimal speech decision \( \eta'_j \), the risk \( r_{ij}(Y(\eta_j)) \) need to be calculated.

1) For speech state \( H^1 \), by substituting the distortion measure \( d_j(X_i, \hat{X}_i) \) of speech-presence into the risk \( r_{ij}(Y(\eta_j)) \), we get
\[
r_{ij}(Y(\eta_j)) = c_2 \int_0^\infty (\lambda_i)' Y_i (X_i - (\lambda_i)' Y_i)' p(X_i | H_i) p(Y(\eta_j)) | X_i | dX_i \] (16)
By solving (16), we can obtain
\[
\hat{A}(Y) = c_2 \lambda_2 \left[ \frac{1}{\lambda_2} \frac{\Gamma(\frac{p+\beta}{2} - 1)}{\phi(\frac{p+\beta}{2} - 1)} \right]^{(\frac{p+\beta}{2} - 1)} \] (17)
where \( \lambda_i = \lambda_2 (k) \eta_i (j = 1, 2) \).

2) For speech state \( H^0 \), by substituting the distortion measure \( d_j(X_i, \hat{X}_i) \) of speech-absent into the risk \( r_{ij}(Y(\eta_j)) \), we get
\[
r_{ij}(Y(\eta_j)) = c_2 \int_0^\infty (\lambda_i)' Y_i (X_i - (\lambda_i)' Y_i)' p(X_i | H_i) p(Y(\eta_j)) | X_i | dX_i \] (18)
By solving (18) based on the Gaussian statistical model [6], we can achieve:
\[
r_{ij}(Y(\eta_j)) = c_2 \lambda_2 (\lambda_i)' Y_i (X_i - (\lambda_i)' Y_i)' \frac{1}{\lambda_2} \exp(-X_i^2) \] (19)
Finally, by substituting \( r_{ij}(Y(\eta_j)) \) and \( r_{ij}(Y(\eta_j)) \) into (9), we can derive the optimal spectral amplitude estimator associated with the optimal speech detector.

3.2. Adaptive calculation of \( p \) and \( \beta \)

From (15), (17) and (19) we can see that, both the optimal speech detector and spectral amplitude estimator depend on the weighted factor \( p \) and power exponent \( \beta \) which values are very important for speech enhancement. However, the \( p \) in [6] and the \( \beta \) in [7] have been calculated according to the overall SNR in each frame such that their values are sole in each frame. And a single value of \( p \) or \( \beta \) is used for all the DFT frequency bins in each frame, which is difficult to obtain a compromise between noise suppression and speech distortion. Therefore, in this paper, the values of \( p \) and \( \beta \) are updated adaptively according to the critical sub-band SNR and the human auditory system (i.e. frequency masking effect and the compressive nonlinearities of the cochlea [8]). That is, the value of \( p \) or \( \beta \) is not equal for different critical bands in each frame. Therefore, a much wider range of gain values can be obtained and then much better performance can be achieved.

3.2.1. Adaptive calculation of weighted factor \( p \)

In this paper, we consider the value of weighted factor \( p \) depends on both the sub-band SNR and the frequency masking threshold in critical band. Then the value of \( p \) can be assumed as a polynomial function [7] of the sub-band SNR \( \Xi(b,k) \) and the masking parameter \( \Theta_{\eta_k}(\lambda, k) \), we have
\[
p(b,k) = \sum_{j=1}^{N} \eta_j b(b,k) \Theta_{\eta_k}(\lambda, k) \] (20)
where \( b \in [0, 21] \) denotes the index of critical band, \( \lambda \) is the frame index, \( k \) is the index of frequency bins, \( \eta_j \) and \( \varphi_i \) (\( i = 0, 1, \ldots, \infty \)) represent the polynomial coefficients, \( \Theta_{\eta_k}(\lambda, k) \) denotes the high-order polynomial terms.

By ignoring the high-order polynomial terms \( \Theta_{\eta_k}(\lambda, k) \), we can approximately rewrite (20) as follows:
\[
p(b,k) = \varphi_1 + \varphi_2 \Xi(b,k) + \varphi_3 \Theta_{\eta_k}(\lambda, k) + \varphi_4 \Xi(b,k) \Theta_{\eta_k}(\lambda, k) + \Theta_{\eta_k}(\lambda, k) \] (21)
where the polynomial coefficients \( \varphi_i \) (\( i = 0, 1, 2, 3 \)) can be obtained through a large number of simulation experiments, which values are 0.542, 0.121, 0.581 and 0.147, respectively. \( \Xi(b,k) \) and \( \Theta_{\eta_k}(\lambda, k) \) are defined as [7]:
\[
\Xi(b,k) = 10 \log_{10} \left( \sum_{l=0}^{L-1} |\tilde{X}(b,k)_l| \right) \] (22)
where \( B_{\text{snr}}(b) \) and \( B_{\text{snr}}(b) \) denote the upper and lower frequency bound of the \( b \)th critical band, respectively. \( \lambda(k) \) is the frequency masking threshold, \( T_{\text{snr}}(\lambda) \) and \( T_{\text{snr}}(\lambda) \) are the minimum and maximum values of the masking threshold in \( \lambda \)th frame, respectively.

3.2.2. Adaptive calculation of power exponent \( \beta \)

According to the cochlea’s compressive nonlinearities [8] [9], we consider the power exponent \( \beta \) as the compression rate of cochlea. Furthermore, the compression rate at high frequencies was measured by some research [8] [9], which is approximately 0.2dB/dB. Thus we set \( \beta_{\text{high}} = 0.2 \) as the lower bound of \( \beta \) value. Although there is no consensus on the degree of nonlinearity at lower frequencies, we can consider the compressive nonlinearities of cochlea are present at lower frequencies and the compression rate is approximately assumed as 0.8dB/dB, which denotes the compression rate at lower frequencies less than it at higher frequencies. Then we choose \( \beta_{\text{low}} = 0.8 \) as the upper bound of \( \beta \) value. For the \( \beta \) value at intermediate frequencies, a linear interpolation method is proposed:

First, from [8] and [9] we can know the fact that, each frequency corresponds to a position on the basilar membrane which called as tonotopic mapping, and the position \( d(k) \) in millimeters is given by:
\[
d(k) = 20 \log_{10}(f(k) + 1) \] (23)
where \( f(k) = k F_c / N \) is the frequency in Hz corresponding to frequency bin \( k \), \( F_c \) is the sampling rate, \( N \) is the frame length, \( A \) is 165 Hz.
Then, we calculate the average position \( d(k,b) \) according to \( d(k) \) in critical band \( b \in (0,21) \), which is defined as follows:

\[
d(b,k) = \frac{1}{B_{b}(k)-B_{b}(k+1)} \sum_{n=0}^{B_{b}(k)} d(k)
\]

(24)

Finally, according to the \( d(k,b) \), the \( \beta \) value can be calculated by linearly interpolating between \( \beta_{hub} \) and \( \beta_{low} \):

\[
\beta(k,b) = \beta_{hub} + \frac{d(k,b) (\beta_{low} - \beta_{hub})}{20 \log_{10} \left( \frac{N}{A} + 1 \right)}
\]

(25)

\[4. \text{ Performance evaluation} \]

The performance evaluation of the proposed method is carried out under the standard of ITU-T G.160 \cite{100}, which mainly includes the amount of noise reduction, the improvement of SNR and the objective quality of enhanced speech.

The clean speech signals are taken from the NTT database. Six types of noise (e.g., White, Babble, Office, Street, Factory, Volvo noise) from ITU-T and NOISEX-92 database are used in our experiments. All speech signals are sampled at 16 kHz and the frame size \( N \) is 512 samples. The samples are hamming windowed with 50% overlap between adjacent frames. In the test, the SDEA in [5] and the weighted Euclidean distortion measure (WEDM) estimator in [6] are used to compare with the proposed algorithm.

\[4.1. \text{ Noise Reduction Test} \]

The test is carried out in white noise, which purpose is to ensure that speech enhancement method can obtain the specified amount of noise reduction \( Q_{a} \) and does not modify speech signal level beyond acceptable range. \( Q_{a} \) and \( Q_{b} \) are the noise reduction factors in the period before and after speech enhancement. The larger \( Q_{a} \) and \( Q_{b} \) are, the more noise attenuation is produced, but their values must be limited in the range of \( Q_{a} \pm 3 \text{dB} \). If the absolute value of \( Q_{a} \) is close to 0, the speech enhancement method will cause less speech distortion. The test results are shown in Table 1.

<table>
<thead>
<tr>
<th>Enhancement method</th>
<th>SNR (dB)</th>
<th>Performance Measurement(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNRI</td>
<td>TNLR</td>
</tr>
<tr>
<td>WEDM 5</td>
<td>13.044</td>
<td>-18.383</td>
</tr>
<tr>
<td>10</td>
<td>11.104</td>
<td>-18.224</td>
</tr>
<tr>
<td>15</td>
<td>9.064</td>
<td>-17.010</td>
</tr>
<tr>
<td>SDEA 5</td>
<td>12.270</td>
<td>-19.147</td>
</tr>
<tr>
<td>10</td>
<td>11.391</td>
<td>-18.172</td>
</tr>
<tr>
<td>15</td>
<td>9.792</td>
<td>-16.331</td>
</tr>
<tr>
<td>Our Method 5</td>
<td>14.065</td>
<td>-19.909</td>
</tr>
<tr>
<td>10</td>
<td>12.223</td>
<td>-18.369</td>
</tr>
<tr>
<td>15</td>
<td>10.228</td>
<td>-17.057</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR Improvement</th>
<th>Enhancement Method</th>
<th>5dB</th>
<th>10dB</th>
<th>15dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEDM 377</td>
<td>0.377</td>
<td>0.303</td>
<td>0.242</td>
<td></td>
</tr>
<tr>
<td>SDEA 343</td>
<td>0.343</td>
<td>0.232</td>
<td>0.209</td>
<td></td>
</tr>
<tr>
<td>Our Method 481</td>
<td>0.481</td>
<td>0.385</td>
<td>0.315</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Objective Quality Test

PESQ improvement is used to assess the improvement of speech quality. The higher PESQ improvement corresponds to better speech quality. The test result is shown in Table 3.

5. Conclusions

We present a single-channel speech enhancement method by coupling speech detection and spectral amplitude estimation. The optimal speech detector and spectral amplitude estimator are derived by jointly minimizing the combined risk function which considers both detection and estimation errors. According to the properties of human auditory perception and compressive nonlinearities of the cochlea, two general weighted cost functions are proposed under speech-presence and speech-absence, respectively. And the cost parameters are employed to balance the speech distortion and residual noise caused by missed detection and false alarm, respectively. We also present the adaptive calculation methods for the weighted factor \( p \) and the power exponent \( \beta \) of cost functions. The test results indicate that the proposed method performs better than the reference methods, which can obtain large amount of noise reduction, and improve the speech quality evidently.

6. Acknowledgements

This work was supported by the Beijing Natural Science Foundation program and Scientific Research Key Program of Beijing Municipal Commission of Education (Grant No. KZ201110005005), the National Natural Science Foundation of China (Grant No. 61072089).
References


