Codebook-Based Speech Enhancement Using Markov Process and Speech-presence Probability

Qi He, Chang-chun Bao and Feng Bao

Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China, 100124
very1990@emails.bjut.edu.cn, baochch@bjut.edu.cn, baofeng@emails.bjut.edu.cn

Abstract

This paper presents a codebook-based speech enhancement algorithm by using Markov process and speech-presence probability (SPP). The Markov process is utilized to model the correlation between the adjacent code-vectors in the codebook for optimizing Bayesian minimum mean squared error (MMSE) estimator. Then the proposed estimator is used to estimate spectral shapes and gains of speech and noise. The correlation between adjacent linear prediction (LP) gains is also fully considered during the procedure of parameter estimation. Through the introduction of SPP in the codebook-constrained Wiener filter, the proposed Wiener filter achieves the goal of much more noise reduction and does not result in the speech component distortion. The evaluation results confirm that the proposed algorithm has a much better performance for reducing annoying background noise than the conventional codebook-based algorithms.

Index Terms: Codebook, Markov process, Wiener filter, Speech enhancement

1. Introduction

Speech enhancement aims to minimize the effect of noise and guarantee both the perceived quality and intelligibility of speech. Speech enhancement algorithm has been of interest for a wide range of applications, such as mobile communications, speech coders, hearing aids and the pre-processing module for robust speech recognition, etc.

The research on single-channel speech enhancement has been conducted over several decades and various algorithms such as spectral subtraction method [1-2], Wiener filtering method [3-4] and subspace based method [5-6] have been proposed. These algorithms perform well in stationary noise environments, but their performance become worse for the non-stationary noises. It is mainly because they adapt too slowly to match the quick variation of noise energy [7-8], which is normal in non-stationary noise environments.

To solve the problem aforementioned, two methods based on the shape codebooks of speech and noise spectra have been developed in [9] and [10], respectively. For these works, the priori knowledge about speech and noise LP parameters are obtained by training and stored in the shape codebooks of speech and noise spectra, respectively. During the enhancement procedure, LP coefficients of speech and noise are estimated by selecting optimal entries from the pre-trained codebooks [9] or computing a weighted sum of the entries [10]. Meanwhile, the spectral gain estimation is achieved adaptively based on the spectral envelope of the observed noisy signal.

Then a Wiener filter constructed from the estimated LP spectra of speech and noise is used to enhance noisy speech. Although the codebook-based methods perform better than conventional methods for non-stationary noise since the spectral gain estimation is online, there are still many annoying residual noises in the enhanced speech, especially in the voiced segments. One reason is that the implied hypothesis about speech-presence is at all time and at all frequencies in conventional codebook-based methods. This is obviously invalid for reality speech, which contains many pauses and may not be present in some particular frequency bins even in the segments of voiced speech. In other word, the SPP almost equals to zero when some noisy frequencies or frames only contain noise. If the codebook-based Wiener filter is modified appropriately by the SPP, we can effectively suppress the noise in these frequency bins or those frames. Furthermore, LP parameters of speech and noise have strong correlation between adjacent frames. Through reasonably utilizing such correlation in the MMSE estimator, we can obtain more accurate estimation of LP parameters.

In this paper, we explicitly consider the correlation between adjacent LP parameters of both speech and noise. And the SPP is introduced to combine with the Wiener filter in the codebook-based enhancement method for the first time.

Fig.1 shows the block diagram of the proposed speech enhancement system. The system consists of two stages: one is the offline training stage which obtains the shape codebooks of speech and noise spectra using LBG algorithm [9]; the other is the real-time de-noising stage. In the stage for de-noising, we get the spectral envelope of the noisy speech by LP analysis (LPA) at first. Then the optimized MMSE estimator is used to obtain the LP parameters of speech and noise according to the pre-trained two codebooks and current spectral envelope of the noisy speech. Thirdly, the posteriori SPP in each frequency bin is calculated using the noisy power spectrum, the spectral envelope estimation of speech and the noise power spectrum obtained by Minima Controlled Recursive Averaging (MCRA) algorithm [11]. Meanwhile, a probability coefficient $\varepsilon$ of current frame is derived from the MCRA algorithm. Fourthly, the Wiener filter constructed from the estimated spectral envelopes of speech and noise is adaptively modified by the posteriori SPP and the probability coefficient $\varepsilon$. Finally, the enhanced speech is obtained by taking the inverse Fast Fourier Transform (IFFT).

The remainder of this paper is organized as follows. In Section 2, we describe the proposed parameter estimator. In Section 3, the structure of modified wiener filter is introduced in details. Our experiments and results are shown in Section 4. Finally, the Section 5 concludes the paper.
2. LP parameters estimation

To formulate the proposed codebook-based speech enhancement system, we define $x(n)$, $s(n)$ and $w(n)$ as the noisy speech, clean speech and noise signals in time frame $n$, respectively. And let $X(k)$, $S(k)$ and $W(k)$ represent the short time Fourier transform (STFT) coefficients of noisy speech, clean speech and noise signals, respectively, at frequency bin $k$, $0 \leq k \leq K-1$ ($K$ denotes the STFT size). Here we consider the additive noise model: $x(n)=s(n)+w(n)$ where speech and noise are assumed to be independent.

2.1. Proposed MMSE estimation of LP parameters

For each pair of code-vectors $[\theta^i, \theta^j]$, where $\theta^i = [\theta_1^i, ..., \theta_p^i]$ and $\theta^j = [\theta_1^j, ..., \theta_p^j]$ are the LP coefficients of speech and noise with $p$ and $q$ orders, respectively, we can obtain the estimated noisy power spectrum as follows:

$$\hat{P}^i(k) = P^i(k) + P^j(k)$$ (1)

where $P^i(k)$ and $P^j(k)$ can be stated as:

$$P^i(k) = g_i(k) / |A_i(k)|^2, \quad P^j(k) = g_j(k) / |A_j(k)|^2$$ (2)

The $1/|A_i(k)|^2$ and $1/|A_j(k)|^2$ represent the spectral shapes related to $\theta^i$ and $\theta^j$, respectively, and $A_i(k)$ and $A_j(k)$ are determined by

$$A_i(k) = 1 + \sum_{n=1}^{N_s} \theta_1^i(n) \exp\left(-\frac{2\pi}{N} nk\right), \quad A_j(k) = 1 + \sum_{n=1}^{N_s} \theta_1^j(n) \exp\left(-\frac{2\pi}{N} nk\right)$$ (3)

The $g_i(k)$ and $g_j(k)$ in (2) are the corresponding LP gains of speech and noise, respectively, which can be obtained by minimizing the Log-spectral (LS) distortion between the actual noisy spectral envelope $P^i(k)$ and the estimated noisy spectrum $\hat{P}^i(k)$. The LS distortion in discrete domain is defined as follows:

$$d_{LS}(P^i(k), \hat{P}^i(k)) = \sum_{k=0}^{N-1} \left| g_i(k) / |A_i(k)|^2 - g_j(k) / |A_j(k)|^2 \right|^2 / \hat{P}^i(k)$$ (4)

By differentiating (4) with respect to the respective gains, and setting the results to zero, the resulting gains can be determined according to the following linear equation:

$$C g_i^* = \mathbf{B}$$ (5)

where the matrices $C$ and $B$ are given in [9].

We define $\theta = [\theta, \theta_\theta, g_s, g_n]$ as the variable representing the speech and noise LP parameters here. Motivated by the high correlation between adjacent LP parameters of both speech and noise, we exploit the previous estimation $\hat{\theta}_{n-1}$ and adjust the MMSE estimation $\hat{\theta}_n$ of $\theta$ in (10) in the $n^{th}$ frame as follows:

$$\hat{\theta}_n = \hat{E}(\theta | x_n, \hat{\theta}_{n-1}) = \hat{f}(\theta | x_n, \hat{\theta}_{n-1}) = \hat{g}_{\theta} \sum_{i=1}^{N_s} \hat{g}^i w^\theta_i$$ (6)

where $p(x_n | \theta_{n-1}, \theta_{\theta_{n-1}}, \theta_{m_{n-1}}, \theta_{m_{n-1}})$ in (6) can be explained as the weighting value of $\theta^i$ in the $n^{th}$ frame. The gains are completely determined by $x_n$, $\theta$, and $\theta_{\theta}$, so we can assume that they are uniformly distribution [10]. The $p(\theta_{n-1}, \theta_{\theta_{n-1}})$, $p(\theta_{m_{n-1}}, \theta_{m_{n-1}})$, $p(g_i | \hat{\theta}_{n-1})$ and $p(g_j | \hat{\theta}_{n-1})$ reflect the inter-frame correlation of LP code-vectors and gains. Next, we will present the method for modeling the correlation in details.

2.2. Modeling correlation of the adjacent code-vectors

LP-coefficient code-vectors of speech and noise have strong correlation between the adjacent frames. In this section, we can use the Markov process [12] to model the correlation between adjacent code-vectors. Each code-vector corresponds to one state of the Markov process. In order to complete a Markov process model for speech, we need to predefine a state transition matrix $a = [a_{ij}]$ (whose dimension is $N_s \times N_s$) and an empirical state transition matrix $b = [b_{ij}]$ (whose dimension is $N_s \times N_s$). The $b_{ij}$ is priori state probability for state $i$ and the $a_{ij} = p(\theta_i | \theta_{i-1})$ is forward transition probability from state $i$ to state $i$. Similarly, the same situations are applied into the noise.
With the $w_i$ defined in (6), we can also obtain the weighting value of speech and noise code-vectors as follows, respectively,

$$w_{s,n} = \sum_{j=1}^{N_w} w_j^s$$

$$w_{n,n} = \sum_{j=1}^{N_w} w_j^n$$

(9)

Figure 2 shows the correlation between the adjacent LP-coefficient code-vectors of speech or noise in Markov process. The $N$ denotes codebook size.

![Figure 2: The correlation between the adjacent code-vectors](image)

According to the inter-frame correlation, the terms $\rho(p_{\theta_s,\theta_n})$ and $\rho(p_{\theta_{s(n-1)},\theta_n})$ in the MMSE estimator can be acquired by:

$$\rho(p_{\theta_{s(n-1)},\theta_n}) = \sum_{j=1}^{N_w} \sum_{j=1}^{N_w} \rho(w_j^s, w_j^n) \rho(p_{\theta_{s(n-1)}}, p_{\theta_n})$$

(10)

$$\rho(p_{\theta_{s(n-1)},\theta_n}) = \sum_{j=1}^{N_w} \sum_{j=1}^{N_w} \rho(w_j^s, w_j^n) \rho(p_{\theta_{s(n-1)}}, p_{\theta_n})$$

(11)

The $\rho(p_{\theta_s})$ and $\rho(p_{\theta_n})$ express the priori probabilities of speech and noise code-vectors, respectively, which obey equal probability distribution [10]. However, this is only valid if an equal number of training vectors fall into each cell [13]. In this paper, the priori probabilities and transition probabilities of states need to be obtained using the LBG algorithm.

### 2.3. Modeling correlation of the adjacent spectral gains

Firstly, the inter-frame log-gain difference $\Delta g_{k}^s$ can be shown as follows:

$$\Delta g_{k}^s = \ln g_{s,k} - \ln g_{s,k-1}$$

(12)

where $g_{s,k}$ and $g_{s,k-1}$ are the LP gains in current and previous frames, respectively. Figure 3 shows the statistical histogram of the speech inter-frame log-gain difference.

![Figure 3: Statistical histogram of $\Delta g_{k}^s$](image)

Obviously, the probability density function (PDF) of $\Delta g_{k}^s$ could be expressed as a zero-mean normal distribution with variance $\psi_{g,k}^2$. The variance $\psi_{g,k}^2$ can be obtained by training. Then given the gain estimation in previous frame, the conditional PDF of LP gain with respect to speech in current frame can be model as a log-normal distribution:

$$p(g_{s,k}^s | \hat{g}_{s,k-1}^s) = \frac{1}{\sqrt{2\pi} \psi_{g,k}^s} \exp \left( \frac{(\ln g_{s,k}^s - \ln \hat{g}_{s,k-1}^s)^2}{2\psi_{g,k}^2} \right)$$

(13)

The modeling method for spectral gain of speech is also applicable for spectral gain of the noise

### 3. Modified Wiener filter based on SPP

After obtaining the estimations of LP parameters, we can apply the following approach to improve the codebook-constrained Wiener filter. We use $H_k(k)$ to denote the event that speech is present at frequency bin $k$. And let $p(H_k(k)) = p(H_k(k))$ denotes the priori probability of speech-presence for frequency bin $k$. Then we can obtain the posteriori SPP shown in Figure 1 as:

$$p(H_k(k) | X(k)) = \frac{\beta(k)}{\beta(k) + (1 - \beta(k))(1 + \xi(k)) \exp(-v(k))}$$

(14)

where $\xi(k) = \xi(k)/(1 + v(k))$ and $\xi(k)$ is the priori signal to noise ratio (SNR) $\xi(k)$ and posteriori SNR $\gamma(k)$ are stated as:

$$\xi(k) = E(p_{\theta_s}(k)) = \frac{\hat{P}_k}{p_{\theta_{mcrc}(k)}} = \frac{\hat{P}_k}{p_{\theta_{mcrc}(k)}}$$

(15)

$$\gamma(k) = \frac{\hat{P}_k}{p_{\theta_{mcrc}(k)}}$$

(16)

The $\hat{P}_k$ is the estimated spectral envelope of speech. The required noise spectrum $p_{\theta_{mcrc}(k)}$ is estimated by the MCRA algorithm. The modified Wiener filter $W_k(k)$ is written as:

$$W_k(k) = \frac{\hat{g}_k}{\hat{A}_k(k)} + \frac{e}{\hat{A}_k(k)}$$

(17)

In order to further attenuate noise in noise-only frame, we get the probability coefficient $e$ from the MCRA algorithm for current frame as follows:

$$e = \frac{1}{K} \sum_{k=0}^{K} p_{\theta_{mcrc}(k)}$$

(18)

where $p_{\theta_{mcrc}(k)}$ is the smoothed SPP for frequency bin $k$ in MCRA algorithm. The final Wiener filter is:

$$W_k(k) = \frac{p(H_k(k) | X(k)) \cdot \hat{g}_k}{\hat{A}_k(k)} + e \cdot \frac{\hat{g}_k}{\hat{A}_k(k)}$$

(19)

Obviously, the $W_k(k)$ can reserve more speech component then $W_k(k)$ for $e=0.5$, i.e., speech component is more than noise component in current frame. When $e=0.5$, i.e. current frame contains more noise component, the $W_k(k)$ can remove more noise than $W_k(k)$. And $W_k(k)$ equals to $W_k(k)$ for $e=0.5$.

### 4. Experiments and evaluations

In this section, the experiments are conducted to verify the effectiveness of the proposed algorithm. The clean speech is
selected from NTT database and down-sampled to 8kHz. The length of each utterance is 8s. Four types of noises from Noisex-92 database are used in our evaluations, including white noise, babble noise, f16 noise, and factory noise. The clean test materials contain ten utterances from three female speakers and two male speakers (two utterances per speaker). And the noisy test materials are created by adding the noises aforementioned onto clean test set at input SNR of 0dB, 5dB, 10dB and 15dB, respectively. The analysis frame is 32ms (256 samples) long, windowed using normalized Hamming window with 50% overlap between the adjacent frames. The STFT size is 512. Tenth-order LP analysis is adopted for both speech and noise. A 6 bit shape codebook of the speech spectrum is trained with one and half hours of clean speech by the LBG algorithm [9]. And the size of shape codebook of the noise spectrum is empirically chosen 3bit, 4bit, 3bit, and 4bit for white, babble, f16, and factory, respectively. The objective measurements used in the evaluations are segmental SNR (SSNR) [14], log-spectral distortion (LSD) [15], and perceptual evaluation of speech quality (PESQ) [16].

The reference methods are the Ref. A [9] method and the Ref. B [10] method. The objective evaluation results are presented in Table 1, Table 2 and Table 3. Each value is an average of ten test utterances under a specific input SNR. CB-WF1, CB-WF2 and CB-WF3 denote the proposed method based on three different modified Wiener filters, respectively. The method of CB-WF1 is the wiener filter method combined with the probability coefficient $\varepsilon$, instead of $p(H(k)X(k))$. The method of CB-WF2 is the wiener filter method combined with the probability $p(H(k)X(k))$, instead of the probability coefficient $\varepsilon$. The method of CB-WF3 is the final wiener filter method with both probability coefficient $\varepsilon$ and $p(H(k)X(k))$.

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR</th>
<th>Ref. A</th>
<th>Ref. B</th>
<th>CB-WF1</th>
<th>CB-WF2</th>
<th>CB-WF3</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0dB</td>
<td>1.48</td>
<td>2.01</td>
<td>2.33</td>
<td>2.26</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>1.75</td>
<td>2.34</td>
<td>2.59</td>
<td>2.59</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>2.14</td>
<td>2.59</td>
<td>2.81</td>
<td>2.84</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>2.47</td>
<td>2.81</td>
<td>3.07</td>
<td>3.08</td>
<td>3.19</td>
</tr>
<tr>
<td>Babble</td>
<td>0dB</td>
<td>1.74</td>
<td>1.77</td>
<td>1.80</td>
<td>1.85</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>2.05</td>
<td>2.16</td>
<td>2.23</td>
<td>2.17</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>2.45</td>
<td>2.46</td>
<td>2.54</td>
<td>2.48</td>
<td>2.53</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>2.80</td>
<td>2.72</td>
<td>2.82</td>
<td>2.75</td>
<td>2.76</td>
</tr>
<tr>
<td>F16</td>
<td>0dB</td>
<td>1.00</td>
<td>2.01</td>
<td>2.04</td>
<td>2.06</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>1.92</td>
<td>2.37</td>
<td>2.45</td>
<td>2.47</td>
<td>2.55</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>2.37</td>
<td>2.68</td>
<td>2.80</td>
<td>2.82</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>2.69</td>
<td>2.93</td>
<td>3.11</td>
<td>3.11</td>
<td>3.16</td>
</tr>
<tr>
<td>Factory</td>
<td>0dB</td>
<td>1.64</td>
<td>1.92</td>
<td>1.99</td>
<td>1.90</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>1.97</td>
<td>2.29</td>
<td>2.38</td>
<td>2.30</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>2.39</td>
<td>2.60</td>
<td>2.71</td>
<td>2.66</td>
<td>2.73</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>2.72</td>
<td>2.85</td>
<td>2.99</td>
<td>2.96</td>
<td>2.98</td>
</tr>
</tbody>
</table>

From the Table 1, we can find that the CB-WF3 leads to higher average PESQ value than the reference methods in all simulated noisy conditions. Comparing with the reference methods and CB-WF1, the CB-WF3 obtains better auditory quality because of the suppression of more background noise. Comparing with the CB-WF2, however, the CB-WF3 does not obtain higher PESQ value at input SNR of 0dB. The main reason of this phenomenon is that the estimation of probability coefficient $\varepsilon$ is not stable enough at low input SNR. In Table 2, the three methods depending on SPP all produced higher average SSNR improvement than the reference methods. And with the decrease of input SNR, the SSNR improvement of the proposed method is much larger than reference methods, which means that the SPP plays an important role for removing the noise especially in silence and unvoiced segments. The test results in Table 3 demonstrate that the CB-WF3 causes lower speech component distortion than reference methods for most of noises. For the white noise, the CB-WF3 produces some higher LSD values than CB-WF2 possibly because of the underestimated value of probability coefficient $\varepsilon$ in some time frames.

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR</th>
<th>Ref. A</th>
<th>Ref. B</th>
<th>CB-WF1</th>
<th>CB-WF2</th>
<th>CB-WF3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5dB</td>
<td>16.55</td>
<td>10.19</td>
<td>7.86</td>
<td>5.64</td>
<td>5.88</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>14.52</td>
<td>8.85</td>
<td>6.00</td>
<td>4.98</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>12.63</td>
<td>7.58</td>
<td>5.48</td>
<td>4.80</td>
<td>4.60</td>
</tr>
<tr>
<td>Babble</td>
<td>0dB</td>
<td>15.48</td>
<td>12.11</td>
<td>11.12</td>
<td>9.99</td>
<td>10.65</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>13.55</td>
<td>10.54</td>
<td>9.57</td>
<td>8.44</td>
<td>9.12</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>11.75</td>
<td>9.07</td>
<td>8.03</td>
<td>7.08</td>
<td>7.74</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>10.07</td>
<td>7.71</td>
<td>6.65</td>
<td>5.99</td>
<td>6.54</td>
</tr>
<tr>
<td>F16</td>
<td>0dB</td>
<td>16.84</td>
<td>11.03</td>
<td>9.69</td>
<td>7.12</td>
<td>8.44</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>14.81</td>
<td>9.53</td>
<td>8.03</td>
<td>5.78</td>
<td>7.11</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>12.91</td>
<td>8.11</td>
<td>6.52</td>
<td>4.98</td>
<td>6.04</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>11.13</td>
<td>6.82</td>
<td>5.30</td>
<td>4.61</td>
<td>5.24</td>
</tr>
<tr>
<td>Factory</td>
<td>0dB</td>
<td>16.79</td>
<td>12.03</td>
<td>11.00</td>
<td>8.27</td>
<td>9.12</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>14.76</td>
<td>10.46</td>
<td>9.41</td>
<td>6.86</td>
<td>7.70</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>12.86</td>
<td>9.00</td>
<td>7.87</td>
<td>5.79</td>
<td>6.55</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>11.09</td>
<td>7.62</td>
<td>6.50</td>
<td>5.09</td>
<td>5.59</td>
</tr>
</tbody>
</table>

5. Conclusions

A codebook-based speech enhancement algorithm based on SPP and Markov process is proposed in this paper. We introduced the Markov process to model the correlation between shape code-vectors of both speech and noise spectra so as to improve the estimation accuracy of LP parameters. And the log-normal distribution is utilized to describe the statistic characteristics of spectral gains of speech and noise, which is better than the conventional Gaussian distribution. Furthermore, the SPP is adopted to incorporate with conventional codebook-based Wiener filter for effectively removing the background noise. In comparison with the reference method, the proposed algorithm could provide better enhanced speech quality and heavier noise reduction in the performance tests.

6. Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61471014).
7. References


