Codebook-based Speech Enhancement with Bayesian LP Parameters Estimation

Qing Wang and Chang-chun Bao
Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China, 100124
E-mail: wangqing221@ emails.bjut.edu.cn and baochch@bjut.edu.cn

***Abstract*—In this paper, we propose a codebook-based Bayesian linear predictive (LP) parameters estimation for speech enhancement, in which the LP parameters are estimated based on the current and past frames of noisy speech. First, by using hidden Markov model (HMM), we develop a new method to drive the speech presence probability (SPP) and speech absence probability (SAP). These two probabilities are the weighting coefficients for the estimated LP parameters corresponding to speech presence and speech absence states. Then we exploit the normalized cross-correction to adjust the transition probabilities between speech-presence and speech-absence states of HMM. The proposed adjustment method makes the SPP estimation more accurately. Finally, in order to suppress the noise between the harmonics of voiced speech, we employ the a posteriori SPP to modify the Wiener filter for enhancing the noisy speech. Our experiments demonstrate that the proposed method is superior to the reference methods.**

***Index Terms*—Speech enhancement, Wiener filter, Speech presence probability, Linear predictive parameters**

I. INTRODUCTION

Speech enhancement in non-stationary noise is still a challenging topic due to its wide applications in hearing aids, mobile communications and speech recognition. Over the past four decades, many speech enhancement algorithms were proposed, such as the typical spectral-subtractive algorithm [1] and Wiener filtering [2]. A major drawback of these algorithms is that their performance will be degraded dramatically when the non-stationary noise is present. In order to overcome this problem, the codebook-based methods [3-6] relying on a priori knowledge about LP coefficients of speech and noise have been proven to work well. However, these methods still retain some artificial noise in the enhanced speech. The main reason is that they derived the LP parameters (LP coefficients and LP gains) on the assumption that speech is always present in noisy signal, but a given segment of noisy observation may consist of noise alone. Due to this unreasonable assumption, the LP parameters estimations of speech and noise are inaccurate. Recently, the researchers proposed a context-based Bayesian speech enhancement technique considering several speech codebooks [5]. Each speech codebook corresponds to its specific hypothesis. The final LP parameters are obtained by weighting the LP parameters based on each hypothesis.

As an example of the codebook-based speech enhancement method, the work in [7] presented a Bayesian framework that considered two hypotheses, i.e., speech presence hypothesis and speech absence hypothesis. Here, the minimum mean-squared error (MMSE) estimation of LP parameters is the weighted sum of the obtained LP parameters under the two hypotheses. The corresponding weighted coefficients are SPP and SAP respectively, and the probabilities vary frame by frame. However, this work [7] does not take full consideration of the inter-frame correlation of voice activity, which results in the inaccuracy estimations of SPP and SAP. In addition, there is another intrinsic problem in the codebook-based speech enhancement methods [3-7]. The priori codebooks only model the spectral envelopes of speech and noise rather than their fine structure, which results in the background noise remained in the voiced segments of the enhanced speech.

In this paper, we propose a codebook-based Bayesian LP parameters estimation for the aforementioned two problems, which estimates the LP parameters based on the information of past noisy speech. First, the HMM theory [8] is employed to derive the SPP and SAP. By using information of past noisy speech, the accuracy of SPP is improved. Then we use the normalized cross-correction coefficient (NCCC) between the spectra of noisy speech and noise [6] to adjust the transition probabilities between speech-presence and speech-absence states of HMM. Finally, the a posteriori SPP in each time-frequency point [9] is combined with the Wiener filter to reduce the noise between the harmonics of the voiced speech.

II. MODEL OVERVIEW

In this section, we provide a brief overview of the memoryless MMSE estimation process in [7]. Considering an additive noise model where clean speech and noise signal are independent, the noisy signal $y(n)$ under the following two hypotheses is given by

$$H_0\text{: speech absent: } y(n) = d(n) \tag{1a}$$

$$H_1\text{: speech present: } y(n) = x(n) + d(n) \tag{1b}$$

where n is the time index, $x(n)$ and $d(n)$ represent the clean speech and noise, respectively. The estimated power spectrum of the noisy signal can be expressed as follows:

$$\hat{P}_y(k) = \hat{P}_x(k) + \hat{P}_d(k) \tag{2}$$

where k is the index of frequency bins. $\hat{P}_x(k) = \sigma_x^2 \big/ \left|A_x(k)\right|^2$ and $\hat{P}_d(k) = \sigma_d^2 \big/ \left|A_d(k)\right|^2$ are the power spectrum estimations of

speech and noise respectively. σ_x^2 and σ_d^2 are the LP gains of respective signals, and

$$A_x(k)=\sum_{l=0}^{p}a_x(l)e^{-jkl},\ A_d(k)=\sum_{l=0}^{q}a_d(l)e^{-jkl} \tag{3}$$

where $\theta_x=(a_x(0),...,a_x(p))$ and $\theta_d=(a_d(0),...,a_d(q))$ denote the LP coefficients of speech and noise, respectively and p, q are the respective LP orders.

We define $m=[m_x,m_d]$ to represent the parameter set of speech and noise. $m_x=[\theta_x,\sigma_x^2]$ is a parameter vector describing the speech power spectrum estimation, and $m_d=[\theta_d,\sigma_d^2]$ describes the noise power spectrum estimation. Under the hypothesis H_0, the conditional expectation of m can be written as:

$$E[m\,|\,\mathbf{y}_n,H_0]=\frac{1}{N_d}\sum_{j=0}^{N_d}m_d^j\frac{p(\mathbf{y}_n|m_d^j)}{p(\mathbf{y}_n\,|\,H_0)} \tag{4}$$

with

$$p(\mathbf{y}_n\,|\,H_0)=\frac{1}{N_d}\sum_{j=0}^{N_d}p(\mathbf{y}_n|m_d^j) \tag{5}$$

Given hypothesis H_1, we have

$$E[m|\mathbf{y}_n,H_1]=\frac{1}{N_xN_d}\sum_{i,j=0}^{N_x,N_d}m^{ij}\frac{p(\mathbf{y}_n|m^{ij})}{p(\mathbf{y}_n|H_1)} \tag{6}$$

with

$$p(\mathbf{y}_n|H_1)=\frac{1}{N_xN_d}\sum_{i,j=0}^{N_x,N_d}p(\mathbf{y}_n|m^{ij}) \tag{7}$$

where N_d and N_x are the codebook sizes of spectral shapes of noise and speech, respectively. Based on the expectation of m given H_0 and H_1, the MMSE estimation of m can be written as:

$$\hat{m}=E[m\,|\,\mathbf{y}]=\sum_{l=0}^{1}E[m\,|\,\mathbf{y},H_l]p(H_l\,|\,\mathbf{y}) \tag{8}$$

where $p(H_l\,|\,\mathbf{y})$ is the SPP or SAP. The estimated parameters m can be used to model the power spectra of speech and noise, respectively. A Wiener filter comprising the estimated power spectral envelops of speech and noise is applied to enhance the noisy speech in the frequency domain, that is,

$$H(k)=\hat{P}_x(k)\big/\hat{P}_x(k)+\hat{P}_d(k) \tag{9}$$

III. Bayesian Estimation

A. Bayesian Estimation of LP Parameters

As explained in section II, the Bayesian framework in [7] only exploits the noisy speech of current frame, which results in the inaccuracy estimation of SPP and SAP. In order to solve this problem, in this section we will develop the estimation methods about SPP and SAP by using the information of current and past frames of noisy speech based on Bayesian framework. Moreover, we exploit the HMM theory to derive SPP and SAP for each noisy frame.

As defined in section II, we consider two hypotheses for the current noisy speech frame. H_0 is speech absence hypothesis and H_1 is speech presence hypothesis. For a given frame n, we define $S=\{S_l=H_l\ ,l=0,1\}$ to denote two states of HMM and q_n denote the HMM state at frame n, i.e., $q_n=S_0$ or $q_n=S_1$.

According to section II, m is the parameter set of speech and noise. The MMSE estimation of m using the information of current and past frames of noisy speech signal can be obtained as follows:

$$\begin{aligned}\hat{m}&=E\left[m|\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n\right]\\&=\sum_{l=0}^{1}E\left[m|\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n,q_n=S_l\right]p(q_n=S_l|\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n)\end{aligned} \tag{10}$$

First, according to the HMM theory [8], we can derive the term $p(q_n=S_l\,|\,\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n)$ as:

$$p(q_n=S_l\,|\,\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n)=\frac{p(\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n,q_n=S_l)}{p(\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n)},\ l=0,1 \tag{11}$$

Let $\alpha_n(l)=p(\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n,q_n=S_l),l=0,1$ represent the forward probability. For the first frame, we have,

$$\alpha_1(l)=p(S_l)p(\mathbf{y}_n\,|\,S_l),\quad l=0,1 \tag{12}$$

where $p(S_0)$ and $p(S_1)$ are the prior probabilities of speech presence and speech absence respectively, and they are assumed to be equal, i.e., $p(S_0)=p(S_1)=0.5$. For the later frames, the forward probability in the current frame could be obtained from each of the forward probability in previous frame with a particular transition probability. i.e.,

$$\alpha_n(l)=\left[\sum_{i=0}^{1}\alpha_{n-1}(i)a_{il}\right]p(\mathbf{y}_n\,|\,S_l),\quad l=0,1 \tag{13}$$

where $a_{il}=p(q_n=S_l|q_{n-1}=S_i)$ is the transition probability, and we assume the transition probability a_{il} to be known beforehand. Since it is well known that speech activities in the adjacent frames have a strong correlation, we set $a_{00}=a_{11}=0.94$ and $a_{01}=a_{10}=0.06$. The term $p(\mathbf{y}_n\,|\,S_l)$ can be obtained by (5) and (7). From (12) and (13), we have

$$p(q_n=S_l\,|\,\mathbf{y}_1,\mathbf{y}_2...,\mathbf{y}_n)=\frac{\alpha_n(l)}{\sum_{i=0}^{1}\alpha_n(i)},l=0,1 \tag{14}$$

The transition probabilities we defined in the preceding way have an inherent problem. The a_{il} taking higher values when $i=l$ may produce time delay and a trailing of estimated SPP at the beginning and the ending segments of speech signal respectively. The trailing phenomenon avoids speech distortion, but the time delay in speech beginning segments would reduce the speech quality. In order to solve this problem, we employ the NCCC [6] to adjust the transition probabilities between speech-presence and speech-absence states of HMM. The NCCC is defined as:

$$\rho=\sum_k(|Y(k)||D(k)|)\Big/\sqrt{\sum_k|Y(k)|^2\sum_k|D(k)|^2} \tag{15}$$

where $|Y(k)|$ and $|D(k)|$ are the amplitude spectra of noisy speech and noise, respectively.

From figure 1, we can see that ρ has a higher value in silence segments, and it decreases to a small value in the voiced segments. In the beginning frame of clean speech, which is indicated with red dashed line, the value of ρ becomes lower than the value in silence segments. Therefore, the value of ρ is applied to adjust the transition probabilities in speech beginning segments, i.e.,

$$a_{\substack{il \\ i,l=0,1}} = \begin{cases} i=l, a_{il}=0.2; i\neq l, a_{il}=0.8 & if\ (p(S_1 \mid \mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_{n-1}) < \delta_1 \\ & and\ \rho_n < \delta_0) \\ i=l, a_{il}=0.94; i\neq l, a_{il}=0.06 & else \end{cases} \quad (16)$$

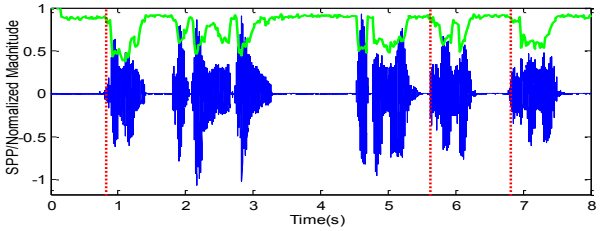

Fig. 1. The example of ρ. The blue solid line is the normalized clean speech waveform, the green solid line is the NCCC, and the red dashed line indicates the speech beginning frame. SNR=5 dB for white noise.

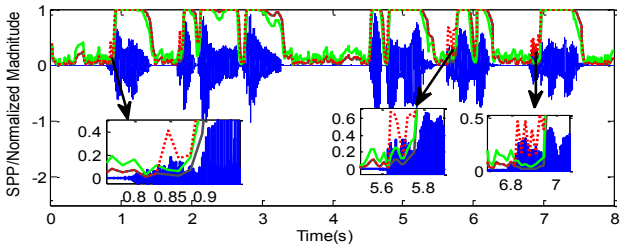

Fig. 2. The example of SPP. The blue solid line is the normalized clean speech waveform. The green solid line is the estimated SPP in [7]. The gray solid line is the proposed SPP estimation without adjusting the transition probabilities, and the red dashed line is the estimated SPP with adjusting the transition probabilities. SNR=5 dB for white noise.

From figure 2, it can be seen that the estimated SPP of proposed method performs better than the method in [7]. The use of NCCC for adjusting the transition probabilities helps the estimate of SPP more accurately between speech-presence and speech-absence states.

Second, we compute the term $E[m \mid \mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_n, q_n = S_n]$. Since the goal of using the past information is to compute SPP and SAP in current frame more accurately, this is obtained by the above work. For a given hypothesis, this term can be derived as in [4]. In terms of the signal non-stationary, the past information given in [4] just contains the current and previous frames noisy speech signal. Here, to retain the focus on the estimation accuracy of SPP and SAP, we assume

$$E[m \mid \mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_n, q_n = S_l] = E[m \mid \mathbf{y}_n, H_l] \quad (17)$$

The way to compute the term $E[m \mid \mathbf{y}_n, H_l]$ is given by the equations of (4), (5), (6) and (7). In order to obtain the estimation of *m*, the equations (14) and (17) are used in (10).

B. Modification of Wiener Filter

As well known, the codebook-based methods [3-7] could not remove the background noise between the harmonics of the enhanced speech in the voiced segments. The main reason is that the priori codebooks cannot model the fine structure of speech and noise spectra. Therefore, we exploit the estimation of a posteriori SPP to modify the Wiener filter in equation (9). The modified Wiener filter can remove the noise between the harmonics in the voiced segments and the noise in silence segments. Here, the a posteriori SPP $\mu(k)$ in [9] is defined as:

$$\mu(k) = \frac{\Lambda(k)}{1+\Lambda(k)} \quad (18)$$

where $\mu(k)$ varies from each time-frequency point, and the generalized likelihood $\Lambda(k)$ is

$$\Lambda(k) = \frac{p(H_k^1)}{1-p(H_k^1)} \left(\frac{1}{1+\bar{\xi}_k}\right)^{\bar{\gamma}_k} \exp\left(\frac{\bar{\xi}_k}{1+\bar{\xi}_k} \frac{\bar{\gamma}_k}{2} \bar{\gamma}_k\right) \quad (19)$$

Using the a posteriori SPP $\mu(k)$, we modify the Wiener filter as:

$$\tilde{I} \qquad k) \frac{\hat{P}_x(k)}{\hat{P}_x(k)+\hat{P}_d(k)} \quad (20)$$

where $\hat{P}_x(k)$ and $\hat{P}_d(k)$ are the power spectrum estimations of speech and noise respectively, which are obtained by using the parameters *m* in (10). $\hat{\mu}(k) = \max\{\mu(k), \mu_{\min}(k)\}$, and $\mu_{\min}(k)$ is a lower bound, which helps to achieve high quality of the enhanced speech by optimizing the trade-off between noise suppression and speech distortion in the voiced segments.

Finally, in order to obtain the enhanced speech, the noisy amplitude spectrum is passed through the modified Wiener filter given in (20).

IV. Experimental Results

In our experiments, one hour of speech utterances are employed to train a 7-bit LSF codebook of speech by LBG algorithm [10]. The test set including twenty speech utterances is selected from NTT database, and the sampling rate of speech signal is 8 kHz. The shape codebooks of noise are trained in a similar way. Four types of noise from NOISE 92 database are used, which include babble, white, office and street. The codebook size of white, babble, street and office are 8, 16, 8 and 8, respectively. The input signal to noise ratio (SNR) is set to 0dB, 5dB and 10dB, respectively. In all cases, speech enhancement is conducted with the experimentally optimized parameter values, $\delta_0 = 0.855$, $\delta_1 = 0.25$, $\mu_{\min} = 0.6$.

For all experiments, the noise codebook is assumed to be known. We consider three references to compare with the proposed algorithms in this paper. Ref. A denotes the codebook-based ML method [3]. Ref. B indicates the codebook-based MMSE method in [4], and Ref. C is the codebook-based MMSE using speech presence uncertainty given in [7]. The average segmental signal-to-noise ratio (SSNR) [11], the average log-spectral distortion (LSD) [12] and the Perceptual Evaluation of Speech Quality (PESQ) [13] are exploited to evaluate the objective quality. In this paper, our evaluation only focuses on the improvement of enhanced amplitude. We therefore exploit the estimated amplitude and the noisy speech phase to reconstruct the enhanced signal.

The table 1 shows average SSNR improvement of the four methods compared to the noisy speech. We find that the proposed method performs better than the three references. In table 2, we show the PESQ test results. Comparing with noisy speech, the PESQ values of these four algorithms all yield greater improvement. In comparison with the three references, The PESQ result of the proposed method still has a much higher value. The test result of LSD is given in table 3. Comparing with noisy speech and the three references, the

proposed method produces a lower LSD value. Thus, since the estimation of SPP is more accurate and the noise reduction between harmonics is more effective, the proposed method achieves a great improvement in terms of the objective quality.

Table 1 Test Result of SSNR Improvement

Noise Type	Method	0dB	5dB	10dB
white	Ref. A	10.33	9.63	8.74
	Ref. B	12.72	11.79	10.98
	Ref. C	19.47	17.78	15.70
	Proposed	22.56	20.13	17.22
babble	Ref. A	6.03	5.30	4.42
	Ref. B	10.25	8.98	7.82
	Ref. C	14.00	12.39	10.82
	Proposed	15.47	12.77	11.92
Street	Ref. A	13.02	11.64	10.17
	Ref. B	16.66	15.10	13.34
	Ref. C	20.15	17.99	15.44
	Proposed	21.64	19.17	16.27
Office	Ref. A	9.73	8.70	7.54
	Ref. B	13.36	12.09	10.64
	Ref. C	16.46	14.81	12.94
	Proposed	17.33	15.60	13.57

Table 2 Test Result of PESQ

Noise Type	Method	0dB	5dB	10dB
white	Noisy	1.38	1.61	1.97
	Ref. A	1.87	2.25	2.53
	Ref. B	2.24	2.49	2.69
	Ref. C	2.23	2.43	2.61
	Proposed	2.34	2.59	2.78
babble	Noisy	1.79	2.13	2.50
	Ref. A	1.81	2.17	2.50
	Ref. B	1.95	2.35	2.66
	Ref. C	1.86	2.22	2.53
	Proposed	1.98	2.38	2.71
Street	Noisy	2.31	2.65	2.95
	Ref. A	2.58	2.85	3.07
	Ref. B	2.79	3.04	3.30
	Ref. C	2.80	3.08	3.36
	Proposed	2.90	3.17	3.43
Office	Noisy	2.01	2.40	2.76
	Ref. A	2.16	2.52	2.81
	Ref. B	2.39	2.73	3.02
	Ref. C	2.39	2.73	3.03
	Proposed	2.46	2.80	3.09

Table 3 Test Result of LSD

Noise Type	Method	0dB	5dB	10dB
white	Noisy	19.12	16.88	15.05
	Ref. A	12.08	10.57	9.17
	Ref. B	9.39	8.08	6.82
	Ref. C	6.55	5.73	5.28
	Proposed	6.56	5.65	5.22
babble	Noisy	14.63	12.61	10.72
	Ref. A	11.22	9.72	8.31
	Ref. B	9.83	8.39	6.99
	Ref. C	8.57	7.09	5.71
	Proposed	7.97	6.51	5.29
Street	Noisy	12.61	10.73	8.99
	Ref. A	8.62	7.23	5.97
	Ref. B	7.37	6.03	4.83
	Ref. C	5.77	4.66	3.87
	Proposed	4.31	4.43	3.97
Office	Noisy	13.09	11.16	9.38
	Ref. A	9.77	8.32	6.97
	Ref. B	8.67	7.22	5.88
	Ref. C	7.47	6.05	4.82
	Proposed	6.11	5.77	4.67

V. CONCLUSIONS

In this paper, Bayesian LP parameters estimation based on the information of current and past noisy speech is proposed for speech enhancement. The HMM theory is employed to drive SPP and SAP. The use of past noisy information and normalized cross-correction makes contribution to improve the accuracy of SPP and SAP. Moreover, by employing the a posteriori SPP to modify the Wiener filter, we can remove the noise between the harmonics of voiced speech more effectively. Our experiments show that the proposed method is superior to the reference methods.

ACKNOWLEDGMENT

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

REFERENCES

[1] Boll, S. F. , Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process., 27(2),113-120. 1979

[2] J. S. Lim and A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proc. IEE, 67(12), 1586-1604, Dec.1979.

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, Codebook driven short-term predictor parameter estimation for speech enhancement, IEEE Trans. Audio, Speech, Lang. Process., vol. 14(1), pp. 163–176, Jan. 2006.

[4] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, Codebook-based Bayesian speech enhancement for nonstationary environments, IEEE Trans. Audio, Speech, Lang. Process., vol. 15(2), pp. 441–452, 2007.

[5] D. H. R . Naidu, S. Srinivasan, Robust Bayesian estimation for context-based speech enhancement,, EURASIP Journal on Audio, Speech, and Music Processing,2014.

[6] F. Bao, H. J. Dou, M. S. Jia and C. C. Bao, Speech enhancement based a few shapes of speech spectrum., IEEE China Summit & Int. Conf. on Signal and Information Processing (ChinaSIP), DOI. 10.1109, pp. 90-94, 2014.

[7] Q. Wang, C. C. Bao, F. Bao, A novel Bayesian framework for speech enhancement using speech presence uncertainty. Chinasip, DOI. 10.1109, pp. 677-681, 2015.

[8] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition., Proc. IEEE. **77**(2), pp. 257–286, 1989

[9] T. Gerkmann, C. Breithaupt, R. Martin, Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. IEEE Trans. Audio, Speech, Lang. Process., vol. 16(5), pp. 910–919, July. 2008.

[10] Y. Linde, A. Buzo and R. Mm Gray. An Algorithm for Vector Quantization Design. IEEE Transactions on Communication, 1980, 28(1):84-95

[11] Quackenbush, S. R., Barnwell, T. P., Clements, M. A., Objective Measures of Speech Quality. Englewood Cliffs, NJ: Prentice Hall, 1988.

[12] Abramson, A., Cohen, I., Simultaneous Detection and Estimation Approach for Speech Enhancement. IEEE Trans. Speech Audio Process., 15(8), pp. 2348-2359, 2007.

[13] ITU-T, Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders, 2001.