SPEECH ENHANCEMENT BASED ON A LOG-SPECTRAL AMPLITUDE ESTIMATOR AND A POSTFILTER DERIVED FROM CLEAN SPEECH CODEBOOK

Jason Wung[1], Biing-Hwang (Fred) Juang[1], and Bowon Lee[2]

[1]Center for Signal and Image Processing, Georgia Institute of Technology
75 Fifth Street NW, Atlanta, GA 30363, USA
{jason.wung, juang}@ece.gatech.edu

[2]Hewlett-Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94304, USA
bowon.lee@hp.com

ABSTRACT

In this paper, we propose a single channel speech enhancement system where a postfilter, which is derived from a clean speech codebook, is applied after a log-spectral amplitude estimator. The primary motivation of this approach is to include prior knowledge about clean source signals to improve speech enhancement results. The codebook, which is trained from clean speech database, serves as clean speech spectral constraints on the enhanced speech. By using the prior clean source information, the proposed method can effectively remove the residual noise presented in traditional speech enhancement algorithms while leaving the speech information intact. Experimental results of the proposed speech enhancement system show improvement in residual noise reduction.

1. INTRODUCTION

The problem of single channel speech enhancement, where the speech signal is corrupted by uncorrelated additive noise, has been widely studied in the past. One of the most popular methods was proposed by Ephraim and Malah [1, 2]. In [1], a short-time spectral amplitude (STSA) estimator is derived from minimum mean square error (MMSE) estimation of the spectral amplitude under the assumption of Gaussian statistical models, where the speech and noise signals are modeled as statistically independent Gaussian random processes. In [2], a log-spectral amplitude (LSA) estimator based on MMSE estimation is also derived. The STSA or LSA estimator is used for the estimation of the short time spectral gain at each frequency bin, where the noisy spectrum is multiplied by the gain to estimate the clean speech spectrum. The gain is a function of the *a priori* signal-to-noise ratio (SNR) and/or the *a posteriori* SNR, where a maximum likelihood (ML) or a “decision-directed” (DD) approach is used for the *a priori* SNR estimation [1]. The LSA estimator is superior to the STSA estimator in that the residual noise level is lowered without increasing the distortion brought upon the noise-reduced speech [2]. However, both the ML and DD SNR estimators cannot completely remove all additive noise and will produce some artifacts in the signal that at times are considered objectionable. The DD SNR estimator leaves colorless residual noise while the ML SNR estimator introduces the annoying “musical noise”. The musical noise is caused by the lack of spectral constraints during spectral amplitude estimation. Without sensible spectral constraints, spectral components in some frequency bins may be unduly boosted or eliminated, resulting in musical noise.

Several methods that may improve the *a priori* SNR estimation have been proposed (*e.g.*, [3–5]). Ren and Johnson [3] estimated the *a priori* SNR from an MMSE estimation perspective, which directly incorporates previous frame information and eliminates the need of empirical weighting factors in the ML and DD SNR estimators Plapous *et al.* [4] estimated the *a priori* SNR in a two-step approach to eliminate the bias introduced by the DD SNR estimator and improve the estimator adaptation speed. Cohen [5] proposed a relaxed statistical model for speech enhancement to take into account the time-correlation between successive speech spectral components for the *a priori* SNR estimation. In these methods, either a Wiener filter [4] or an LSA estimator [3, 5] is used as the spectral gain function.

All of the approaches mentioned above rely on the accuracy of the *a priori* SNR estimation to lower the residual noise level, without directly addressing the removal of residual noise. To address the residual noise issue, a codebook-based postfiltering method [6] was proposed recently, where a postfilter was applied after the LSA estimator. The postfilter is constructed based on a combination of prototypical clean speech spectra, which are obtained *a priori* from clean speech through vector quantization or Gaussian mixture modeling. The postfilter aims at reducing the residual noise or artifacts so as to make the final result most resembling a clean speech signal in terms of statistical characteristics. The spectral constraints take advantage of the frequency dependencies which are not considered in traditional speech enhancement algorithms, where the spectral component in each frequency bin is independently estimated. By imposing the spectral constraints, the spectral peaks of the noisy signal can be further enhanced. In the meantime, the artifacts can be reduced.

In [6], the postfilter consists of a weighted sum of the model spectra derived from the codebook, where the postfilter weights are obtained based on the likelihood ratio distortion. However, the processed speech sounds muffled with this approach. Since the weighted sum of the model spectra incorporates all codewords, it is equivalent to applying a filter that effectively averages those codewords to one instance of spectrum. This is effectively applying an averaged speech spectrum, which has a spectral roll-off at high frequency. In this paper, we derive alternative solutions to the postfilter weights that are mathematically more tractable and alleviate the muffledness issue. Specifically, postfilter weights based on MMSE and non-negative least squares (NNLS) are discussed.

The paper is organized as follows. In Section 2, we review the LSA estimator with ML and DD *a priori* SNR estimation approaches. In Section 3, we present the codebook-based postfilter. Enhancement results are presented in Section 4 and conclusion is given in Section 5.

2. MMSE LOG-SPECTRAL AMPLITUDE ESTIMATION

Let $x[n] \equiv x(nT)$ and $d[n] \equiv d(nT)$ denote the clean speech and noise samples, respectively, where T is the sampling period and n is the sample index. Let $y[n] \equiv y(nT)$ denote the noisy speech samples, which is given by

$$y[n] = x[n] + d[n].$$

Let $Y_k(m) \equiv R_k(m)e^{j\phi_k(m)}$, $X_k(m) \equiv A_k(m)e^{j\theta_k(m)}$, and $D_k(m) \equiv N_k(m)e^{j\psi_k(m)}$ be the k^{th} spectral component, in the m^{th}

analysis window, of the noisy signal $y[n]$, the clean speech signal $x[n]$, and the noise $d[n]$, respectively.

The objective is to find an estimator $\hat{X}_k(m)$ which minimizes the conditional expectation of a distortion measure given a set of noisy spectral measurements. Let $\mathbf{Y}_k(m') \equiv \{Y_k(m'), Y_k(m'-1), \ldots, Y_k(m'-L+1)\}$ denote a set of L spectral measurements and $d(X_k(m), \hat{X}_k(m))$ denote a given distortion measure between $X_k(m)$ and $\hat{X}_k(m)$. Therefore, $\hat{X}_k(m)$ can be estimated as [5]

$$\hat{X}_k(m) = \arg\min_X \mathcal{E}\left\{ d(X_k(m), X) \,\middle|\, \mathbf{Y}_k(m') \right\},$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator.

Without loss of generality, assuming that the current frame is m, we define the log spectral amplitude distortion

$$d_{\text{LSA}}(X_k, \hat{X}_k) \equiv |\log A_k - \log \hat{A}_k|^2. \tag{1}$$

Under the assumption of Gaussian statistical model, where the speech and noise are modeled as statistically independent complex Gaussian random variables with zero mean, an estimate for X_k is obtained by applying a spectral gain function to the noisy spectral measurements

$$\hat{X}_k = G(\xi_k, \gamma_k) Y_k,$$

where the *a priori* and *a posteriori* SNRs are defined as

$$\begin{aligned} \xi_k &\equiv \lambda_X(k)/\lambda_D(k), && \textit{a priori} \text{ SNR}, \\ \gamma_k &\equiv |Y_k|^2/\lambda_D(k), && \textit{a posteriori} \text{ SNR}. \end{aligned}$$

$\lambda_X(k) \equiv \mathcal{E}\{|X_k|^2\}$ and $\lambda_D(k) \equiv \mathcal{E}\{|D_k|^2\}$ denote the variances of the k^{th} spectral components of the clean speech and the noise, respectively. Using (1), the gain function is given by [2]

$$G_{\text{LSA}}(\xi_k, \gamma_k) = \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2} \int_{\nu_k}^{\infty} \frac{e^{-t}}{t} \, \mathrm{d}t \right),$$

where ν_k is defined by

$$\nu_k \equiv \frac{\xi_k}{1+\xi_k} \gamma_k.$$

Therefore, we need to estimate the *a priori* SNR ξ_k as well as the noise variance $\lambda_D(k)$. Note that the estimation of noise variance is not the focus in this paper. It can be estimated by using methods such as minimum statistics [7] or minima controlled recursive averaging [8].

2.1 Decision-Directed Estimation

The DD *a priori* SNR estimation is given by [1]

$$\hat{\xi}_k^{\text{DD}}(m) = \alpha \frac{|\hat{X}_k(m-1)|^2}{\lambda_D(k, m-1)} + (1-\alpha) \mathrm{P}\{\gamma_k(m) - 1\},$$

where $\hat{X}_k(m-1)$ is the amplitude estimate of the k^{th} signal spectral component in the $(m-1)^{\text{th}}$ analysis frame, $\alpha \in [0,1]$ is a weighting factor, and $\mathrm{P}\{\cdot\}$ is defined as

$$\mathrm{P}\{x\} \equiv \begin{cases} x, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The name "decision-directed" comes from the fact that the *a priori* SNR is updated based on the previous frame's amplitude estimation.

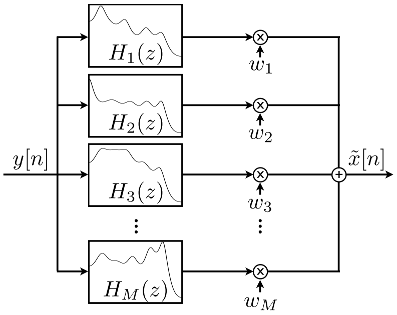

Figure 1: A block diagram of the proposed postfiltering model.

2.2 Maximum Likelihood Estimation

The ML estimation is based on estimation of signal variance by maximizing the joint conditional probability density function (PDF) of $\mathbf{Y}_k(m)$ given $\lambda_X(k)$ and $\lambda_D(k)$, which can be written as

$$\hat{\lambda}_X^{\text{ML}}(k) = \arg\max_{\lambda_X(k)} \left\{ p(\mathbf{Y}_k(m) \,\middle|\, \lambda_X(k), \lambda_D(k)) \right\}.$$

This estimator results in the following *a priori* SNR estimator

$$\hat{\xi}_k^{\text{ML}}(m) = \begin{cases} \frac{1}{L} \sum_{l=0}^{L-1} \gamma_k(m-l) - 1, & \text{if non-negative,} \\ 0, & \text{otherwise,} \end{cases}$$

where estimation is based on L consecutive frames $\mathbf{Y}_k(m) \equiv \{Y_k(m), Y_k(m-1), ..., Y_k(m-L+1)\}$, which are assumed to be statistically independent. The actual implementation is a recursive average given by [1]

$$\begin{aligned} \bar{\gamma}_k(m) &= \alpha \bar{\gamma}_k(m-1) + (1-\alpha) \frac{\gamma_k(m)}{\beta}, \\ \hat{\xi}_k^{\text{ML}}(m) &= \mathrm{P}\{\bar{\gamma}_k(m) - 1\}, \end{aligned}$$

where $\alpha \in [0,1]$ and $\beta \geq 1$ are both weighting factors.

3. THE PROPOSED POSTFILTER

Prototypical clean speech spectra are obtained from a clean speech database through codebook training. Postfiltering is done by passing the noisy speech signal or the LSA enhanced speech signal through a postfilter $H(z)$, which is given by

$$H(z) \equiv \sum_{i=1}^{M} w_i H_i(z),$$

where M is the number of codewords, $H_i(e^{j\omega}) = 1/A_i(e^{j\omega})$ is the frequency response of an all-pole filter corresponding to the model spectrum derived from the i^{th} codeword based on linear prediction (LP) analysis, and w_i is the postfilter weight of the i^{th} filter. A block diagram of this model is shown in Figure 1. Without loss of generality, we can drop the frame index m and define the postfiltered spectral estimate at each frequency bin k as

$$\tilde{X}_k \equiv Y_k H(k) = Y_k \sum_{i=1}^{M} w_i H_i(k). \tag{2}$$

The name "postfilter" comes from the fact that the postfilter weights are obtained after the LSA enhancement step. Two possible ways of obtaining the postfilter weights are discussed below.

3.1 Postfilter Weights Based on the MMSE Criterion

(2) can be reformulated as

$$\tilde{\mathbf{x}} = \mathbf{C}\mathbf{w},$$

where $\tilde{\mathbf{x}} = [\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_K]^{\mathrm{T}}$, $\mathbf{w} = [w_1, w_2, \ldots, w_M]^{\mathrm{T}}$, and $\mathbf{C}$ is a matrix where the j^{th} column vector is given by

$$\mathbf{c}_j = \begin{bmatrix} Y_1 H_j(1) \\ Y_2 H_j(2) \\ \vdots \\ Y_K H_j(K) \end{bmatrix}, \quad \forall j \in 1, 2, \ldots, M.$$

Deriving the postfilter weights based on the MMSE criterion leads to the following optimization problem

$$\hat{\mathbf{w}}^{\text{MMSE}} = \arg\min_{\mathbf{w}} \mathcal{E}\left\{ \|\mathbf{x} - \mathbf{C}\mathbf{w}\|^2 \right\}. \tag{3}$$

The estimation error is defined as

$$e = \|\mathbf{x} - \mathbf{C}\mathbf{w}\|^2 = \sum_{k=1}^{K} |X_k - \tilde{X}_k|^2,$$

where K is the total number of frequency bins. The minimum value of $\mathcal{E}\{e\}$ occurs when the gradient is zero. Evaluating the gradient and we have

$$\begin{aligned} \frac{\partial \mathcal{E}\{e\}}{\partial w_j} &= \frac{\partial}{\partial w_j} \sum_{k=1}^{K} \mathcal{E}\left\{ |\tilde{X}_k|^2 \right\} - \mathcal{E}\left\{ 2\Re\{X_k^* \tilde{X}_k\} \right\} \\ &= 2\sum_{i=1}^{M} w_i \sum_{k=1}^{K} H_i(k) H_j(k) \mathcal{E}\left\{ |Y_k|^2 \right\} \\ &\quad - 2\sum_{k=1}^{K} H_j(k) \mathcal{E}\left\{ \Re\{X_k^* Y_k\} \right\} \\ &= 0, \quad \forall j \in 1, 2, \ldots, M, \end{aligned} \tag{4}$$

where $\Re\{\cdot\}$ denotes the real value. Under the assumption of additive noise model and that the noise and speech are independent Gaussian random variables with zero mean, we have $\mathcal{E}\{|Y_k|^2\} = \lambda_X(k) + \lambda_D(k)$ and $\mathcal{E}\{\Re\{X_k^* Y_k\}\} = \lambda_X(k)$. After Substituting the above terms into (4), we have

$$\sum_{i=1}^{M} w_i \sum_{k=1}^{K} H_i(k) H_j(k) [\lambda_X(k) + \lambda_D(k)] = \sum_{k=1}^{K} H_j(k) \lambda_X(k),$$

which can be rewritten as a system of equations $\mathbf{T}\mathbf{w} = \mathbf{b}$, where $\mathbf{T}$ is a matrix with each element given by

$$\begin{aligned} t_{ij} &= t_{ji}, \quad \forall i, j \in 1, 2, \ldots, M, \\ t_{ji} &= \sum_{k=1}^{K} H_i(k) H_j(k) [\lambda_X(k) + \lambda_D(k)]. \end{aligned}$$

t_{ij} is element in the i^{th} row and j^{th} column of matrix $\mathbf{T}$, and $\mathbf{b} = [b_1, b_2, \ldots, b_M]^{\mathrm{T}}$, where

$$b_j = \sum_{k=1}^{K} H_j(k) \lambda_X(k), \quad \forall j \in 1, 2, \ldots, M.$$

Therefore, we can use the output of speech enhancement algorithms to estimate $\lambda_X(k)$ and use a noise variance estimate for $\lambda_D(k)$. In our experiments, $\lambda_X(k)$ for the MMSE postfilter is estimated as

$$\hat{\lambda}_X(k) = |\hat{X}_k^{\text{LSA}}|^2 \equiv |G_{\text{LSA}}(\xi_k, \gamma_k) Y_k|^2, \tag{5}$$

where ξ_k comes from either the ML or the DD estimation. The optimal postfilter weights can be determined by solving $\mathbf{w} = \mathbf{T}^{-1}\mathbf{b}$. Since the postfilter weights obtained from the MMSE criterion can result in negative values, the overall spectral gain function is chosen as

$$\tilde{X}_k^{\text{MMSE}} \equiv Y_k \left| \sum_{i=1}^{M} \hat{w}_i^{\text{MMSE}} H_i(k) \right|.$$

3.2 Postfilter Weights Based on Non-negative Least Squares

Non-negativity constraints on the postfilter weights can be imposed by reformulating (3) as an NNLS problem

$$\begin{aligned} \hat{\mathbf{w}}^{\text{NNLS}} = \arg\min_{\mathbf{w}} \|\mathbf{x} - \mathbf{C}\mathbf{w}\|^2, \quad & \text{subject to } w_i \geq 0, \\ & \forall i \in 1, 2, \ldots, M. \end{aligned} \tag{6}$$

By using NNLS to limit the solution space of the postfilter weights, most of the postfilter weights will be zero in a given frame. Therefore, zero weights are assigned to the spectral prototypes which deviate from the spectral shape of the speech spectrum in that frame. On the other hand, if the NNLS postfilter is applied to the noisy speech, only the overall background noise level will be reduced while the noise between speech harmonics will be retained. Therefore, the NNLS postfilter is applied after the LSA filtered signal to suppress the residual noise of the LSA filtered speech

$$\tilde{X}_k^{\text{NNLS}} \equiv \hat{X}_k^{\text{LSA}} \sum_{i=1}^{M} \hat{w}_i^{\text{NNLS}} H_i(k).$$

In our actual implementation, the following is used to solve (6)

$$\begin{aligned} \mathbf{x} &= [\lambda_X(1), \lambda_X(2), \ldots, \lambda_X(K)]^{\mathrm{T}}, \\ \mathbf{c}_j &= \begin{bmatrix} [\lambda_X(1) + \rho(1)\lambda_D(1)] H_j(1) \\ [\lambda_X(2) + \rho(2)\lambda_D(1)] H_j(2) \\ \vdots \\ [\lambda_X(K) + \rho(k)\lambda_D(K)] H_j(K) \end{bmatrix}, \quad \forall j \in 1, 2, \ldots, M, \end{aligned}$$

where $\lambda_X(k)$ is given by (5) and $\rho(k) \in [0, 1]$ is an attenuation factor which is determined by the residual noise level. The reason for this modification is that we are reducing only the residual noise from the LSA filtered speech rather than all the noise from the noisy speech. For low SNR bins, $\rho(k)$ has to be small to prevent over attenuation of the residual noise, while for high SNR bins, the value of $\rho(k)$ does not have great impact since $\lambda_X(k) \gg \rho(k)\lambda_D(k)$. For this reason, we choose $\rho(k) = G_{\text{LSA}}(\xi_k, \gamma_k)$.

4. EXPERIMENTAL RESULTS

Experiments to evaluate the proposed algorithm were performed using the TIMIT database. The sampling frequency is 16 kHz. A frame size of 512 samples with 75% overlap was used. A Hamming window was applied on each frame during training and testing. Codebook training was performed using 4620 sentences of clean speech and testing was performed using 9 noisy speech utterances. The speech database for testing were different from those used for training. Both male and female speakers were included. The codebook was trained with truncated cepstral distance distortion measure. A 24th order LP analysis was used and the order of truncated cepstral coefficients was 48. These parameters are different from those in [6] due to different sampling frequencies. Gaussian white noise, F16 cockpit noise, and babble noise were added to each testing utterance at segmental signal-to-noise ratio (SSNR) of -5, 0, 5, and 10 dB. Both the DD and the ML *a priori* SNR estimation were used for the LSA filter. For the DD estimation, the weighting factor was $\alpha = 0.98$, whereas the weighting factors were $\alpha = 0.725$ and $\beta = 2$ for the ML estimation. The speech variance

Table 1: SSNR improvement for Gaussian white noise.

Input	LSA-	ML-	ML-	LSA-	DD-	DD-
SSNR	ML	NNLS	MMSE	DD	NNLS	MMSE
-5 dB	8.01	8.93	**9.09**	7.04	8.66	**8.79**
0 dB	6.29	7.27	**7.37**	5.63	7.16	**7.44**
5 dB	4.79	5.72	**5.83**	4.22	5.56	**5.96**
10 dB	3.51	4.38	**4.49**	3.03	4.15	**4.63**

Table 2: SSNR improvement for F16 cockpit noise.

Input	LSA-	ML-	ML-	LSA-	DD-	DD-
SSNR	ML	NNLS	MMSE	DD	NNLS	MMSE
-5 dB	7.29	8.04	**8.22**	6.27	7.65	**7.85**
0 dB	5.56	6.45	**6.56**	4.87	6.29	**6.61**
5 dB	4.11	5.04	**5.07**	3.59	4.91	**5.26**
10 dB	2.99	**3.93**	3.91	2.58	3.80	**4.10**

Table 3: SSNR improvement for babble noise.

Input	LSA-	ML-	ML-	LSA-	DD-	DD-
SSNR	ML	NNLS	MMSE	DD	NNLS	MMSE
-5 dB	6.60	**7.79**	7.74	6.26	7.75	**7.86**
0 dB	4.88	5.98	**6.24**	4.78	6.11	**6.42**
5 dB	3.51	4.65	**4.73**	3.42	4.72	**5.12**
10 dB	2.45	**3.55**	3.55	2.37	3.62	**3.94**

Table 4: LSD for Gaussian white noise.

Input	LSA-	ML-	ML-	LSA-	DD-	DD-
SSNR	ML	NNLS	MMSE	DD	NNLS	MMSE
-5 dB	5.32	**4.80**	4.90	5.27	**4.98**	5.01
0 dB	3.94	**3.56**	3.59	3.99	3.77	**3.74**
5 dB	2.72	2.49	**2.44**	3.01	2.74	**2.58**
10 dB	1.74	1.63	**1.53**	2.07	1.85	**1.63**

Table 5: LSD for F16 cockpit noise.

Input	LSA-	ML-	ML-	LSA-	DD-	DD-
SSNR	ML	NNLS	MMSE	DD	NNLS	MMSE
-5 dB	5.07	**4.71**	4.71	4.95	4.93	**4.84**
0 dB	3.67	**3.32**	3.37	3.74	3.55	**3.49**
5 dB	2.53	**2.27**	2.29	2.74	2.47	**2.39**
10 dB	1.64	**1.44**	1.45	1.87	1.59	**1.50**

Table 6: LSD for babble noise.

Input	LSA-	ML-	ML-	LSA-	DD-	DD-
SSNR	ML	NNLS	MMSE	DD	NNLS	MMSE
-5 dB	5.03	**4.63**	4.67	4.64	4.71	**4.64**
0 dB	3.51	**3.14**	3.20	3.39	3.29	**3.21**
5 dB	2.42	**2.08**	2.15	2.46	2.19	**2.16**
10 dB	1.58	**1.33**	1.37	1.71	1.40	**1.37**

estimates for the MMSE postfilter and the NNLS postfilter were obtained from the LSA filtered speech. The noise variance estimate was obtained by recursively averaging past spectral power values of the noise

$$\hat{\lambda}_D(k,m) = \eta\hat{\lambda}_D(k,m-1) + (1-\eta)|D_k(m)|^2,$$

where $\eta = 0.85$.

The MMSE postfilter results were based on a codebook size of 128, while the NNLS postfilter results were based on a codebook size of 1024. If the codebook size of the MMSE postfilter is too large, the inverse problem $\mathbf{w} = \mathbf{T}^{-1}\mathbf{b}$ can become ill-conditioned. Therefore, a relatively smaller codebook size for the MMSE postfilter is chosen. On the other hand, the NNLS postfilter does not have this constraint and a larger codebook size provides finer resolution for the codeword selection, at the expense of longer computation.

Two objective measurements were chosen for evaluation: SSNR and log spectral distortion (LSD), which and are defined as [5]

$$\text{SSNR} = \frac{1}{J}\sum_{m=0}^{J-1}\mathcal{T}\left\{10\log_{10}\frac{\sum_{n=0}^{N-1}x^2[n+\frac{Nm}{4}]}{\sum_{n=0}^{N-1}(x[n+\frac{Nm}{4}]-\hat{x}[n+\frac{Nm}{4}])^2}\right\},$$

$$\text{LSD} = \frac{1}{J}\sum_{m=0}^{J-1}\left\{\frac{1}{\frac{K}{2}+1}\sum_{k=0}^{K/2}\left[10\log_{10}\frac{\mathcal{C}X_k(m)}{\mathcal{C}\hat{X}_k(m)}\right]^2\right\}^{\frac{1}{2}},$$

where J is the number of frames, $N = 512$ is the size of a frame, $\mathcal{T}$ confines the SNR at each frame to perceptually meaningful range between 35 dB and -10 dB, *i.e.,* $\mathcal{T}\{x\} \equiv \min\{\max\{x,-10\},35\}$, and $\mathcal{C}X_k(m) \equiv \max\{|X_k(m)|^2,\delta\}$ is the clipped spectral power such that the log-spectrum dynamic range is confined to 50 dB, where $\delta \equiv 10^{-50/10}\max_{k,m}\{|X_k(m)|^2\}$.

For simplicity, let LSA-DD and LSA-ML denote the LSA filters using the DD and the ML *a priori* SNR estimation, respectively. ML-MMSE and ML-NNLS denote the MMSE and the NNLS postfilters based on LSA-ML output, while DD-MMSE and DD-NNLS denote the MMSE and the NNLS postfilters based on LSA-DD output. Table 1, 2, and 3 show the results of SSNR improvement using LSA filter, NNLS postfilter, and MMSE postfilter. The MMSE postfilter shows the highest improvement most of the time, while the performance of the NNLS postfilter closely follows. Applying the postfilter always improve SSNR results. Table 4, 5, and 6 show the LSD for all enhancement algorithms. In most cases, the postfilters yield lower LSD than the LSA filters.

Figure 2 shows the spectrogram of clean, noisy, LSA filtered speech, and postfiltered speech in their respective panels, where the noise type is Gaussian white noise with 5 dB input SSNR. The LSA-ML filter has a higher output SSNR than the LSA-DD filter at the expense of musical noise, which can be attributed to isolated frequency spikes in high frequency area. On the other hand, the residual noise level of the LSA-DD filter is still quite high compared to LSA-ML. The postfilter removes both the musical noise of the LSA-ML filter as well as the residual white noise of the LSA-DD filter. MMSE postfilter performs more aggressively than the NNLS postfilter in terms of the removal of residual noise, which can also be verified by the SSNR improvement in Table 1, 2, and 3.

A subjective listening study shows that the proposed method can successfully remove most of the residual noise from the LSA filtered speech. Both the MMSE and NNLS postfiltered speech provides much lower residual noise level than the LSA filtered speech. Even though the objective scores such as SSNR and LSD are better on the MMSE postfiltered speech, the NNLS postfiltered speech sounds more naturally pleasing, since the MMSE postfiltered speech may sound too clean and unnatural. On the other hand, small amount of residual noise from the LSA filtered speech can still be perceived in the NNLS postfiltered speech, which can also be observed from Figure 2.

5. CONCLUSION

A speech enhancement system based on a codebook driven postfilter was discussed in the paper. Since the codebook is derived from a clean speech database, it imposes spectral constraints on either the noisy speech signal or the LSA filtered signal. The postfilter consists of a weighted sum of the codeword, where the postfilter weights are derived from MMSE and NNLS methods. Experimental results show that the postfilter can effectively remove the residual noise of the LSA filters. Objective measurements based on SSNR and LSD also confirm the improved speech enhancement results.

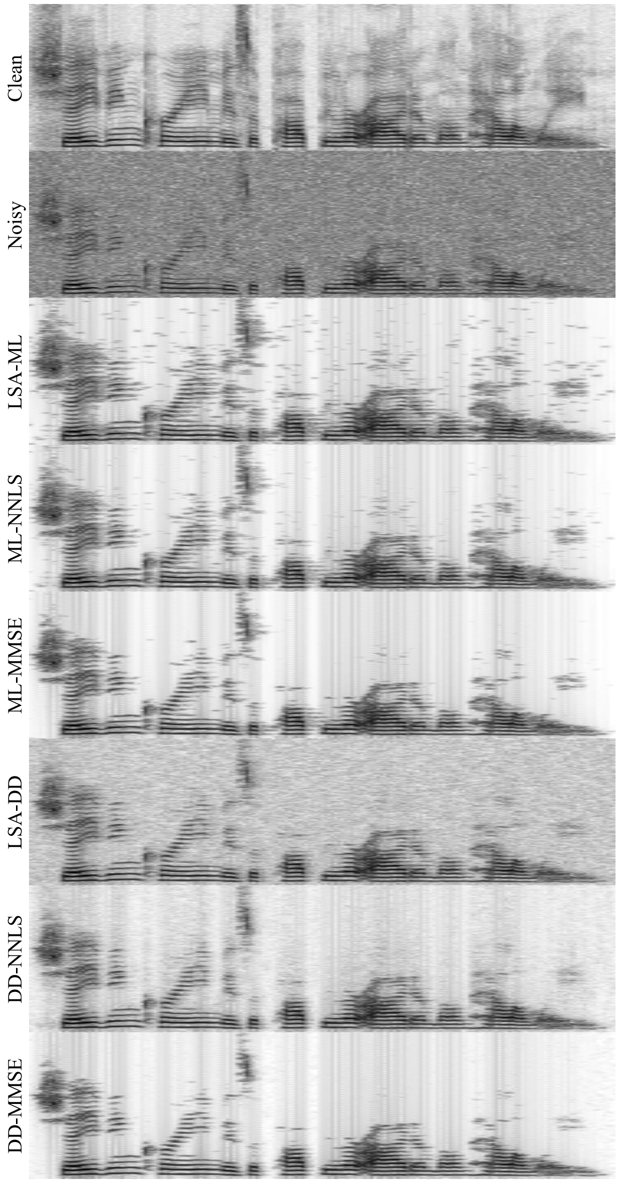

Figure 2: Spectrograms of clean speech, Gaussian white noise corrupted speech, and enhanced speech at 5 dB input SSNR

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109 – 1121, Jan 1984.

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443 – 445, Jan 1985.

[3] Y. Ren and M. Johnson, "An improved SNR estimator for speech enhancement," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4901 – 4904, Jan 2008.

[4] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2098 – 2108, 2006.

[5] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 870 – 881, 2005.

[6] J. Wung, S. Miyabe, and B.-H. Juang;, "Speech enhancement using minimum mean-square error estimation and a post-filter derived from vector quantization of clean speech," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4657 – 4660, 2009.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504 – 512, Jul 2001.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 466 – 475, Sep 2003.