Proceeding of the IEEE 28th
Canadian Conference on Electrical and Computer Engineering
Halifax, Canada, May 3-6, 2015

On the Use of a Codebook-based Modeling Approach for Bayesian STSA Speech Enhancement

Golnaz Ghodoosipour, Eric Plourde and Benoit Champagne

***Abstract*— In this paper, we develop a Bayesian short-time spectral amplitude (STSA) estimator with the purpose of single-channel speech enhancement in the presence of moderate levels of non-stationary noise. In this regard, we first apply a minimum mean squared error (MMSE) approach for the joint estimation of the short-term predictor (STP) parameters of the speech and noise signals, from the noisy speech observations. This approach is based on using trained codebooks of speech and noise linear predictive (LP) coefficients to model the *a priori* information needed by the MMSE estimation. Afterwards, the power spectra derived from the estimated STP are passed to the Wβ-SA STSA estimator, where they are used to calculate the enhancement gains applied to the short-term Fourier transform (STFT) coefficients of the noisy speech. When compared to an existing benchmark approach from the literature, the proposed approach combining codebook-based STP estimation with the Wβ-SA method gives rise to a notable improvement in the quality of the processed noisy speech.**

***Index Terms*— speech enhancement, Bayesian STSA estimation, codebook methods, linear prediction**

I. INTRODUCTION

Speech enhancement aims to remove background disturbances from a noisy speech signal while preserving the naturalness and intelligibility of the processed output. Over the years, a wide range of processing algorithms operating under different conditions have been proposed for this task [1]. Among these, a specific class of frequency domain, single channel methods, collectively known as Bayesian STSA estimators, have received considerable attention due to their superior performance and low complexity [2]. In these methods, the estimate of the clean speech is obtained by applying a gain to the STFT coefficients of the noisy speech, where the gain in each frequency bin is derived by minimizing the expected value of a cost function which provides a measure for the error in the clean speech STSA estimate.

In the well-known MMSE STSA estimator, the cost function is chosen as the mean squared error between the estimated and the true speech STSA, whose minimization leads to a closed form expression for the optimal processing gain under the Gaussian assumption [3]. In the β-order STSA MMSE estimator (denoted as β-SA, for short), the β-th power of the STSA is taken prior to computing the squared error in the cost function. Such non-linear compression can be justified based on the characteristics of human hearing and therefore leads to a more perceptually relevant criterion [4]. In the weighted Euclidean STSA MMSE estimator (denoted as WE-SA), the error spectrum is weighted by the inverse of the clean speech STSA raised to power α [5]. This approach, which is based on the perceptually weighted error criterion in speech coding, also improves the enhancement performance. A generalization of these methods called Wβ-SA is proposed in [6], where both a weighting factor and a power law, chosen based on the characteristics of the human auditory system, are used in the definition of the cost function. Consequently, the resulting STSA estimator is shown to achieve better enhancement performance than its counterparts.

In all these methods, the enhancement made to the noisy speech depends on statistical properties of the desired speech and the corrupting noise. In particular, the variance of the speech and noise components as functions of the frequency, referred to as power spectral density (PSD) in this context, must be estimated as part of the enhancement process. This estimation is particularly problematic for the corrupting noise, specially when the latter is non-stationary and its statistics change over time. Several traditional approaches are available for noise power estimation, including the use of voice activity detection [7], minimum tracking [8] and recursive averaging [9], but their use in connection with Bayesian STSA estimators in the presence of non-stationary noise does not necessarily provide the expected level of performance.

In past years, sophisticated methods have been developed in which data-driven statistical learning is applied to obtain *a priori* knowledge of the speech and noise descriptors. This knowledge is captured to develop probabilistic models of the observed data which, in turn, can be employed to derive estimators of the relevant speech and noise statistics. As an example, in [10], the parameters of the speech and noise spectral shapes, specifically the auto-regressive (AR) coefficients and associated excitation gains, are modeled using hidden Markov models (HMM). Other examples of such model based systems, are the methods which use trained codebooks of speech and noise LP coefficients [11]-[12]. In contrast to HMM based methods, which include the excitation gains in the *a priori* information, the gains in [12] are assumed to be unknown and estimated along with the LP coefficients directly from the observed noisy speech using a Bayesian formulation derived from a codebook model.

Support for this work was provided by a CRD grant from NSERC (Govt. of Canada) and Microsemi (Ottawa, Canada).

G. Ghodoosipour and B. Champagne are with the Department of Electrical and Computer Engineering, McGill University, 3480 University St., Montreal, Canada, H3A 0E9 (emails: golnaz.ghodoosipour@mail.mcgill.ca; benoit.champagne@mcgill.ca).

E. Plourde is with the Département de génie électrique et de génie informatique, Université de Sherbrooke, 2500 boul. de l'Université, Sherbrooke, Canada, J1K 2R1 (email: eric.plourde@usherbrooke.ca).

978-1-4799-5829-0/15/$31.00 ©2015 IEEE

In this paper, we investigate the incorporation of the codebook-based method for spectral parameter estimation within the Bayesian STSA speech enhancement framework, where special emphasis is given to the Wβ-SA estimator. The specific codebook-based estimator used here is a combination of the methods proposed in [11] and [12]. At first, the maximum likelihood (ML) estimates of the speech and noise excitation gains are derived using the method proposed in [11]. Then the ML estimates are used to obtain Bayesian MMSE estimators of the speech and noise LP coefficients, along with updated excitation gains, using the method in [12]. The resulting spectral parameters are employed to derive the final signal and noise PSDs needed in the computation of the optimal enhancement gain, as per the Wβ-SA approach. Since the estimate of the noise PSD is constantly updated, this method performs efficiently in non-stationary environments. In particular, its performance is compared to that of the STFT-based Wiener filtering method [13]. The results point to the superiority of the Wβ-SA estimator over the Wiener filter when used in combination with codebook-based PSD estimation.

This paper is organized as follows. In Section II, the Wβ-SA speech enhancement method is described, while in Section III, the codebook based method for PSD estimation is reviewed. In Section IV, the incorporation of these two methods is explained in detail, followed by the presentation and discussion of objective evaluation results in Section V. Section VI contains a brief concluding statement.

II. Wβ-SA SPEECH ENHANCEMENT METHOD

We consider frame based processing, with the observed noisy speech in a particular frame ℓ given by

$$y_\ell(n) = x_\ell(n) + w_\ell(n), \qquad 0 \le n < N \tag{1}$$

where $y_\ell(n)$, $x_\ell(n)$ and $w_\ell(n)$ denote the samples of the noisy speech, the desired speech and the additive noise, respectively, integer n is the discrete-time index and N is the frame length. Following the application of the STFT with proper analysis window, we let $Y_{k,\ell}$, $X_{k,\ell}$, and $W_{k,\ell}$ refer to the k^{th} complex spectral components of the noisy speech, clean speech, and noise in the ℓ^{th} frame, respectively. In the sequel, to simplify the notations, the frame index ℓ will be omitted, unless otherwise indicated.

In the Bayesian STSA framework of speech enhancement, the main goal is to obtain an estimator $\hat{\mathcal{X}}$ of $\hat{\mathcal{X}}_k \triangleq |X_k|$, i.e. the STSA of the clean speech signal. The desired estimate can be derived by minimizing the expected value of a cost function which measures the error between $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$, as given by:

$$\hat{\mathcal{X}}_k = \arg\min_{\hat{\mathcal{X}}_k} E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\} \tag{2}$$

where E denotes statistical expectation. The resulting STSA estimate is then combined with the phase of the noisy speech, i.e.:

$$\hat{X}_k = \hat{\mathcal{X}}_k e^{j\angle Y_k} \tag{3}$$

to obtain an estimate of the complex spectrum of the clean speech. The latter is inverse Fourier transformed and pass to an overlap-add module in order to reconstruct the enhanced speech in the time-domain.

The problem of estimating the clean speech with such Bayesian estimator mainly depends on defining a proper cost function $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$ and associated statistical models for the signal and noise components. In effect, several different Bayesian STSA estimators have been developed in this way. In early work [3], it is proposed to use the squared error between the clean speech STSA and its estimate as the cost function, that is:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2. \tag{4}$$

Then, by modeling the STFT coefficients of the speech and noise as independent circular complex Gaussian random variables, a closed form solution is obtained for the optimal estimator in (2). The resulting solution is generally known as the MMSE-STSA estimator in the literature.

Subsequently, other Bayesian estimators were proposed by generalizing the MMSE STSA method. In the β-SA estimator [4], the β-th power of the estimated and clean speech STSA is taken prior to computing the squared error in the cost function (4), which is replaced by $C = (\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^{\ \beta})^2$. In the WE-SA estimator, the squared error is weighted by the pth power of the clean speech STSA, i.e., $C = \mathcal{X}^p(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$. These different cost functions all lead to closed-form solutions when used in connection with (2), thereby offering different trade-offs between noise reduction and speech distortion.

In [6], a new family of Bayesian STSA estimators, referred to as Wβ-SA, is proposed where the cost function includes both a power law and a weighting factor

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta}{\mathcal{X}_k^\alpha} \right)^2 \tag{5}$$

where α is related to the parameter p in the WE-SA estimator [5] through $\alpha = -p/2$ and β is related to the β-SA estimator [4]. In the Wβ-SA estimator, these parameters are chosen based on the human auditory system and ear's masking properties; as a result they become frequency dependent. This characteristic of the Wβ-SA estimator results in a better noise reduction while controlling the speech distortion.

Using the cost function (5), along with the Gaussian statistical model in [3] for the signal and noise components, the optimization problem (2) can be solved in closed-form. The resulting estimator can be expressed as:

$$\hat{\mathcal{X}}_k = G_k |Y_k| \tag{6}$$

where $G_k > 0$ is the gain applied to the spectral magnitude of the noisy speech. Specifically, it is shown in [6] that the gain for the Wβ-SA estimator takes the form:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left(\frac{\Gamma\left(\frac{\beta}{2} - \alpha + 1\right) M\left(\alpha - \frac{\beta}{2}, 1; -\upsilon_k\right)}{\Gamma(-\alpha + 1)\, M(\alpha, 1; -\upsilon_k)} \right)^{1/\beta} \tag{7}$$

where $\Gamma(a)$ stands for the gamma function and $M(a,b;c)$ is the confluent hypergeometric function. The other gain parameters in [6] are defined as

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \qquad \xi_k = \frac{\sigma^2_{X,k}}{\sigma^2_{W,k}}, \qquad \gamma_k = \frac{|Y_k|^2}{\sigma^2_{W,k}} \tag{8}$$

where $\sigma^2_{X,k} = E\{\mathcal{X}_k^2\}$ and $\sigma^2_{W,k} = E\{|W_k|^2\}$ denote the speech and noise variances, respectively. In the literature, the parameters ξ_k and γ_k are referred to as the *a priori* and *a posteriori* SNRs, respectively.

III. CODEBOOK-BASED ESTIMATION OF SPEECH AND NOISE SPECTRAL PARAMETERS

In the proposed approach, ML estimates of the speech and noise excitation gains are first obtained using trained codebooks of speech and noise LP coefficients, such as in [11]. These estimates of the excitation variances are then employed to obtain MMSE estimators of the speech and noise LP coefficients, as per [12]. These estimation approaches are briefly reviewed below.

In [11], a codebook based method is proposed wherein the gains of speech and noise are assumed to be deterministic, but unknown. The codebook, which is obtained with training data, provides a model for the underlying distribution of the speech and noise LP coefficients. An ML estimation approach is then applied to determine the unknown excitation gains of the speech and noise component by searching through the codebook for the maximum of the likelihood function.

Let $\boldsymbol{\theta}_x^i$ and $\boldsymbol{\theta}_w^j$ represent the ith and jth codebook vectors of speech and noise LP coefficients, respectively:

$$\boldsymbol{\theta}_x^i = [a^i_{x_0}, \cdots, a^i_{x_p}], \quad \boldsymbol{\theta}_w^j = [a^j_{w_0}, \cdots, a^j_{w_q}] \tag{9}$$

where p and q are the LP model orders. The ML estimates of the speech and noise excitation variances are obtained by using these codebooks according to:

$$\{\sigma_x^{2,ML}, \sigma_w^{2,ML}\} = \arg \max_{i,j,\sigma_x^2,\sigma_w^2} p_y(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j; \sigma_x^2, \sigma_w^2) \tag{10}$$

where $\boldsymbol{y} = [y(0), y(1), ...y(N-1)]$ is the vector of noisy speech observations in the given frame, and where σ_x^2 and σ_w^2 refer to the speech and noise excitation variances. The likelihood function can be written as [11]:

$$p_y(\boldsymbol{y}|\boldsymbol{\theta}_x, \boldsymbol{\theta}_w; \sigma_x^2, \sigma_w^2) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{R}_y|^{1/2}} e^{-\frac{1}{2}\boldsymbol{y}^T \boldsymbol{R}_y^{-1}\boldsymbol{y}} \tag{11}$$

where $\mathbf{R}_y = E\{\mathbf{y}\mathbf{y}^T\}$ is the covariance matrix of the noisy speech, which can be expressed as $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_w$. In turn, the speech and noise covariance matrices, $\mathbf{R}_x$ and $\mathbf{R}_w$ respectively, can be expressed in terms of the model parameters. For instance,

$$\mathbf{R}_x = \sigma_x^2(\mathbf{A}_x^T\mathbf{A}_x)^{-1} \tag{12}$$

where $\mathbf{A}_x$ is the $N \times N$ lower triangular Toeplitz matrix with $[1, a_{x_1}, ..., a_{x_p}, 0, ...0]^T$ as the first column.

In [11], using some analysis and numerical approximations, it is shown that the ML estimate of the excitation gains in (10) can be obtained by solving the following linear equations:

$$C \begin{bmatrix} \sigma_x^{2,ML} \\ \sigma_w^{2,ML} \end{bmatrix} = D \tag{13}$$

$$C = \begin{bmatrix} \big\||A_w^j(\omega)|^4\big\| & \big\||A_x^i(\omega)|^2|A_w^j(\omega)|^2\big\| \\ \big\||A_x^i(\omega)|^2|A_w^j(\omega)|^2\big\| & \big\||A_x^i(\omega)|^4\big\| \end{bmatrix} \tag{14}$$

$$D = \begin{bmatrix} \big\|P_y(\omega)|A_x^i(\omega)|^2|A_w^j(\omega)|^4\big\| \\ \big\|P_y(\omega)|A_x^i(\omega)|^4|A_w^j(\omega)|^2\big\| \end{bmatrix} \tag{15}$$

where $P_y(\omega) = |Y(\omega)|^2$ is the observed noisy power spectrum, and $A_x^i(\omega)$ and $A_w^j(\omega)$ denote the AR spectra derived from the LC coefficients as follows:

$$A_x^i(\omega) = \sum_{k=0}^{p} a^i_{x_k} e^{-j\omega k}, \quad A_w^j(\omega) = \sum_{k=0}^{p} a^j_{w_k} e^{-j\omega k} \tag{16}$$

In our work, we will use the ML excitation gains derived as above to obtain the final MMSE estimators of the speech and noise LP coefficients by applying the approach in [12]. The LP coefficients along with the excitation gain are generally referred to as the short term predictor parameters (STP). The complete set of STP parameters for the speech and noise signals can be represented by a single vector, i.e.,

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_x, \boldsymbol{\theta}_w, \sigma_x^2, \sigma_w^2] \tag{17}$$

which is modeled as a random vector with joint probability density function (PDF) $p(\boldsymbol{\theta})$. The main goal here is to estimate $\boldsymbol{\theta}$ given the noisy speech observations contained in vector $\mathbf{y}$.

In the MMSE approach introduced in [12], an estimator of $\boldsymbol{\theta}$ which minimizes the mean square error $E\{(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta})^2\}$ is derived. The solution to this problem is well known to be the conditional expectation, that is:

$$\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\boldsymbol{y}) = E\{\boldsymbol{\theta}|\boldsymbol{y}\}. \tag{18}$$

Expanding the expected value, we can rewrite (18) as follows:

$$\hat{\boldsymbol{\theta}} = \int_\Theta \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = \int_\Theta \boldsymbol{\theta} \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} d\boldsymbol{\theta} \tag{19}$$

where $\Theta = \Theta_x \times \Theta_w \times \Sigma_x \times \Sigma_w$, Θ_x and Θ_w represent the support-spaces of the vectors of LP coefficients of speech and noise, and Σ_x and Σ_w are the support-spaces for the speech and noise excitation gains.

While the conditional PDF $p(\mathbf{y}|\theta)$ is available from (11), in order to evaluate (19), expressions for the PDFs of vectors $\boldsymbol{\theta}$ and $\mathbf{y}$ are needed. In [12], the former is obtained by assuming that $\boldsymbol{\theta}_x$, $\boldsymbol{\theta}_w$, σ_x^2 and σ_w^2 are mutually independent, and that the excitation gain are uniformly distributed, which leads to

$$p(\boldsymbol{\theta}) = c\ p(\boldsymbol{\theta}_x)p(\boldsymbol{\theta}_w) \tag{20}$$

where c is a normalization constant.

Furthermore, it is argued that the conditional PDF $p(\mathbf{y}|\theta)$ is concentrated around the ML values of the excitation gains as the latter are varied. That is, it is implicitly assumed that for a test function $\phi(\boldsymbol{\theta})$, we have $\int\int \phi(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\sigma_x^2 d\sigma_w^2 \approx$c'$\phi(\boldsymbol{\theta}')p(\boldsymbol{y}|\boldsymbol{\theta}')$ where c' is a constant and $\boldsymbol{\theta}' = [\boldsymbol{\theta}_x, \boldsymbol{\theta}_w, \sigma_x^{2,ML}, \sigma_w^{2,ML}]$. Under this approximation, the PDF of $\boldsymbol{y}$ can be obtained as shown in (21) at the bottom of this page.

In [12], assuming that the codebook has high-dimensionality, integrals over the LP parameter space $\Theta_x \times \Theta_w$ with measure $p(\boldsymbol{\theta}_x)p(\boldsymbol{\theta}_w)\boldsymbol{\theta}_x\boldsymbol{\theta}_w$ are approximated by discrete sums over the corresponding codebook entries. Specifically, let us introduce $\boldsymbol{\theta}'_{i,j} = [\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML}]$ where $\boldsymbol{\theta}_x^i$ and $\boldsymbol{\theta}_w^j$ are the ith and jth codebook vectors of speech and noise LP coefficients, and $\sigma_{x,ij}^{2,ML}$ and $\sigma_{w,ij}^{2,ML}$ are the ML estimates of the speech and noise excitation gains based on $\boldsymbol{\theta}_x^i$ and $\boldsymbol{\theta}_w^j$ as given by (13). Under the above discrete approximation, the integrals in (19) and (21) can be approximated by the summation over the codebook entries shown in (22) and (23) at the bootom of this page, where N_x and N_y are the speech and noise codebook sizes, respectively.

IV. COMBINATION OF SPEECH ENHANCEMENT AND CODEBOOK BASED ESTIMATION

As it can be observed from (7), the optimal gain G_k in the Wβ-SA method is a function of the *a posteriori* SNR, γ_k, and of v_k, where in turn v_k is a function of γ_k and the *a priori* SNR, ξ_k, as defined in (8). Therefore, it can be concluded that only γ_k and ξ_k need to be evaluated to compute the clean speech estimate. These two parameters are dependent on the noise and speech variances denoted by $\sigma_{X,k}^2$ and $\sigma_{W,k}^2$.

In order to combine the Wβ-SA method of speech enhancement with the codebook based estimation approach, we first propose to replace the speech and noise variance parameters $\sigma_{X,k}^2$ and $\sigma_{W,k}^2$ in (8) by the corresponding LP-based PSD estimates obtained from the codebook method. Specifically, for each frequency bin k, we propose the following substitution:

$$\sigma_{X,k}^2 \rightarrow \hat{P}_x(\omega_k) \triangleq \frac{\hat{\sigma}_{x,k}^2}{|\hat{A}_x(\omega_k)|^2} \tag{24}$$

$$\sigma_{W,k}^2 \rightarrow \hat{P}_w(\omega_k) \triangleq \frac{\hat{\sigma}_{w,k}^2}{|\hat{A}_w(\omega_k)|^2} \tag{25}$$

where the spectra $\hat{A}_x(\omega)$ and $\hat{A}_w(\omega)$ are defined as in (16), but using the codebook-based MMSE estimates of the speech and noise LP coefficients, i.e. $\hat{\boldsymbol{\theta}}_x$ and $\hat{\boldsymbol{\theta}}_w$, and where $\hat{\sigma}_{x,k}^2$ and $\hat{\sigma}_{w,k}^2$ are the corresponding MMSE gain estimates. That is, following initial generation of the speech and noise codebooks with training data, for each frame of time-domain observations $\boldsymbol{y}$, a noisy power spectrum $P_y(\omega_k) = |Y(\omega)|^2$ is computed and used to derive the codebook-based estimator $\hat{\boldsymbol{\theta}}$ in (22). The resulting optimum parameters are used in turns to compute the required speech and noise variances as in (24)-(25).

The *a priori* SNR ξ_k is defined in (8) as the ratio of the clean speech variance to that of the noise. Under the independence assumption for the speech signal and noise, the relationship between the *a priori* and *a posteriori* SNR can be expressed as:

$$\xi_k = E\{\gamma_k - 1\}. \tag{26}$$

Combining the two equations i.e. (8) and (26), we can obtain a recursive estimator of ξ_k at the ℓ-th frame, denoted by $\xi(k,\ell)$, via the following operation:

$$\xi(k,\ell) = \tau \frac{G(k,\ell-1)^\rho |Y(k,\ell-1)|}{\sigma_W^2(k,\ell-1)} + (1-\tau)\frac{\sigma_X^2(k,\ell)}{\sigma_W^2(k,\ell)} \tag{27}$$

where τ is a weighting or smoothing factor in the range of $0.95 \leq \gamma < 1$, and $\sigma_X^2(k,\ell)$ and $\sigma_W^2(k,\ell)$ are the ℓ^{th} speech and noise PSDs obtained from (24) and (25) for the ℓth frame. The parameter ρ in (27) offers a trade-off between noise reduction and distortion; experimentally, we find that a value of $\rho < 1$ produces better results. This nonlinear smoothing has the advantage of eliminating large power variations in consecutive frames, which in turn tends to reduce the level of musical noise [14].

V. RESULTS

In this Section, we investigate the performance of the above proposed speech enhancement method by presenting the results of selected enhancement experiments for different types of speech and noise backgrounds. At first, the quality of the enhanced speech obtained with the proposed algorithm is examined. Afterwards, we compare the new

$$p(\boldsymbol{y}) = \int_{\Theta_x}\int_{\Theta_w} p(\boldsymbol{y}|\boldsymbol{\theta}_x, \boldsymbol{\theta}_w, \sigma_x^{2,ML}, \sigma_w^{2,ML})p(\boldsymbol{\theta}_x)p(\boldsymbol{\theta}_w)d\theta_x d\theta_w \tag{21}$$

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_x N_w}\sum_{i=1}^{N_x}\sum_{j=1}^{N_w} \boldsymbol{\theta}'_{i,j} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML})p(\boldsymbol{\theta}_x^i)p(\boldsymbol{\theta}_w^j)}{p(\boldsymbol{y})} \tag{22}$$

$$p(\boldsymbol{y}) = \frac{1}{N_x N_w}\sum_{i=1}^{N_x}\sum_{j=1}^{N_w} p(\boldsymbol{y}|\boldsymbol{\theta}_x^i, \boldsymbol{\theta}_w^j, \sigma_{x,ij}^{2,ML}, \sigma_{w,ij}^{2,ML})p(\boldsymbol{\theta}_x^i)p(\boldsymbol{\theta}_w^j) \tag{23}$$

method, which combines the Wβ-SA speech enhancement with the codebook-based approach for estimating the noise and speech statistics, with a similar but alternative scheme where the Wβ-SA is replaced by the Wiener filter [13].

An 8-bit speech codebook of LP coefficients of dimension $p = 10$ is trained using the generalized Lloyd algorithm (GLA) [15]. The speech training set consists of 4 minutes of recorded clean speech from 2 male and 2 female speakers, available from the McGill TSP database [16]. A 3-bit noise codebook of LP coefficients of dimension $q = 10$ is trained in the same way. The noise training set consists of a 10 minutes concatenation of five different types of recorded noise from the AURORA database [17], that is: car, train, street, restaurant and airport noise. Selected noise segments (that were not part of the training set) are added to the clean speech and enhancement experiments are conducted for noisy speech with input SNR of 0, 5, 10 and 15 dB. The sampling frequency of the speech and noise signals is set to F_s = 8 kHz. In the application of the STFT, a frame length of $N = 256$ samples with 50% overlap is used, along with the Hanning window. In the STFT domain, the enhancement of the noisy speech is carried out by applying a gain to the noisy speech STFT, as in (6). The processed frames are inverse transformed back to the time-domain where they are reassembled with the overlap-add method.

We compare two different methods, namely the combination of Wβ-SA with codebook based estimation of spectral statistics, as described in Section IV, and the combination of Wiener filter [13], with codebook based approach. For the Wiener filter, the enhancement gain is computed as:

$$G_k = \frac{1}{1 + \dfrac{\sigma^2_{w,k}|A_x(\omega_k)|^2}{\sigma^2_{x,k}|A_w(\omega_k)|^2}} \tag{28}$$

The two algorithms are compared in terms of the ITU-T recommended objective measure, i.e. perceptual evaluation of speech quality (PESQ).

A. Enhanced speech waveforms

The time-domain signal waveforms of the noisy, true and the enhanced speech for a selected experiment are plotted in Fig. 1 in order to demonstrate the algorithm's effectiveness. The results correspond to the speech of a male speaker, contaminated by the street noise at SNR = 5 dB. As it can be observed, the proposed algorithm performs relatively well in removing street noise from the noisy speech.

B. Objective PESQ results

In Table I the average PESQ objective measures of three speakers in different environments including train, airport, car and street noise are presented. The results are given for three different values of SNR, i.e. 0 dB, 5 dB, and 10 dB. According to the results, the proposed algorithm, which combines the Wβ-SA estimator with the codebook-based approach for the speech and noise estimation, is superior to the Wiener filter based combination in all cases in terms of the PESQ measure. Except for the case of street noise at high SNR, the improvement in PESQ with the use of the Wβ-SA method are significant. These results are also consistent with informal listening tests.

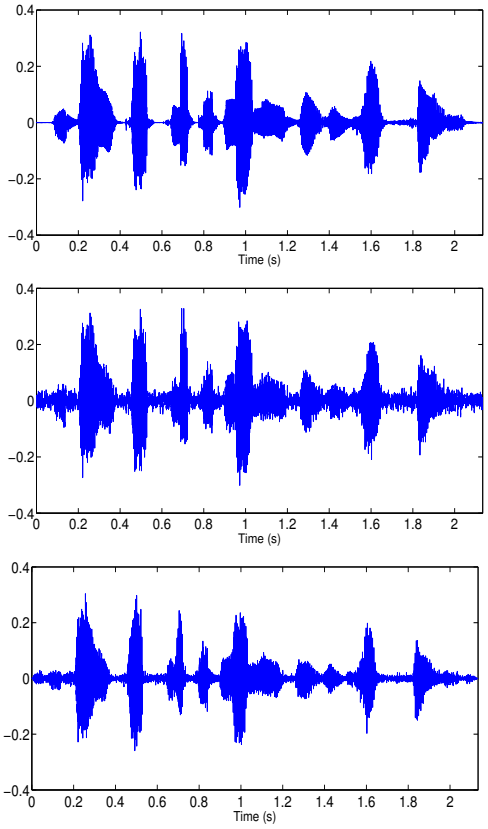

Fig. 1. Time domain waveforms, for a female speaker and train noise at SNR=10dB. From top to bottom: clean speech, noisy speech, enhanced speech

TABLE I

AVERAGE OF PESQ OBJECTIVE MEASURES FOR ENHANCEMENT OF NOISY SPEECH FROM THREE MALE AND FEMALE SPEAKERS

Noise type	Method	SNR 0 dB	SNR 5 dB	SNR 10 dB
Train noise	Wiener	1.61	1.87	2.18
	Wβ-SA	1.89	2.29	2.51
Airport noise	Wiener	1.61	2.14	2.31
	Wβ-SA	1.96	2.28	2.48
Street noise	Wiener	1.03	1.36	2.34
	Wβ-SA	1.23	1.78	2.37
Car noise	Wiener	1.54	1.90	2.22
	Wβ-SA	1.86	2.23	2.45

VI. CONCLUSION

In this paper, we developed a Bayesian STSA estimator for the purpose speech enhancement in the presence of non-stationary noise. To this end, trained codebooks of speech and noise LP coefficients were used to model the required *a priori* information needed for STSA estimation. The codebooks were employed to derive MMSE estimators of the speech and noise STP parameters, which in turn were used in combination with the Wβ-SA method to obtain the final estimator of the clean speech. It was shown that the

proposed scheme performs well in terms of removing a significant amount of noise from the noisy speech. It was also shown that when compared to the combination of the codebook-based method with the traditional Wiener filter, the proposed speech enhancement approach gave rise to a notable improvement of the quality of the processed noisy speech in terms of the PESQ objective measure.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[2] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms", *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 153-156, Vol. 1, May 2006.

[3] Y. Ephraim and D. Malah , "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoustics, Speech, Signal Process.*, pp. 1109-1121, Vol. 32, Dec. 1984.

[4] C. H. You, S. N. Koh and S. Rahardja , "β-order MMSE spectral amplitude estimation for speech enhancement", *IEEE Trans. Speech, Audio Process.*, pp. 475-486, Vol. 4, Jul. 2005.

[5] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum", *IEEE Trans. Speech, Audio Process.*, pp. 857-869, Vol. 13, Aug. 2005.

[6] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement", *IEEE Trans. Audio, Speech, Language Process.*, pp. 1614-1623, Vol. 16, Nov. 2008.

[7] A. Sangwan , W. Zhu and M. Ahmad, "Improved voice activity detection via contextual information and noise suppression", *IEEE Int. Symp Circuits Systems (ISCAS)*, pp. 868-871, May 2005.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Trans. Speech, Audio Process.*, pp. 504-512, Vol. 9, 2001.

[9] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *IEEE Signal Process. Lett.*, pp. 12-15, Vol. 9, Jan. 2002.

[10] H. Sameti, L. Deng, H. Sheikhzadeh and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise", *IEEE Trans. Speech, Audio Process.*, pp. 445-455, Vol. 6, Sept. 1998.

[11] M. Kuropatwinski and W. B. Kleijn , "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding", *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pp. 669-672, Vol. 1, May 2001.

[12] S. Srinivasan, J. Samuelsson and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments", *IEEE Trans. Audio, Speech, Language Process.*, pp. 441-452, Vol. 2, Feb. 2001.

[13] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, Wiley-Teubner, Englewood Cliffs, 1998.

[14] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", *IEEE Trans. Audio, Speech, Process.*, pp. 345-349, Vol. 2, 1994.

[15] Y. Linde , A. Buzo and R. M. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Trans. Communications*, pp. 84-95, Vol. 28, Jan. 1980.

[16] P. Kabal, "*TSP Speech Database*", *Telecommunications and Signal Processing Laboratory*, Mcgill University, 2002, http://www-mmsp.ece.mcgill.ca/Documents/Data/index.html.

[17] H. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition", *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, 2000, France, http://aurora.hsnr.de.