STOCHASTIC INTEGRATION AND LONG-TERM PREDICTOR ESTIMATION UNDER NOISY CONDITIONS FOR SPEECH ENHANCEMENT

M. Kuropatwinski and W.B. Kleijn

KTH (Royal Institute of Technology), Signal, Sensors and Systems Dept.
Stockholm, Sweden

ABSTRACT

We propose a method to estimate the short term predictor (STP) and the long-term predictor (LTP) under noisy conditions. We assume the speech signal to be a single, dual or triple frame asymptotic mean stationary process. The *a priori* STP parameter distribution is represented as databases sampled from the speech training data. Stochastic integration is used to obtain the minimum mean square error estimates of the STP parameters. After computing the STP parameters, the LTP parameters from a database of pairs of taps and excitation variances are matched, together with the lag, using a likelihood criterion, to the noisy speech. The estimated STP and LTP parameters are also applied to obtain clean speech estimates by means of a Wiener or a Kalman filter. For car noise with an SNR of -5dB, the proposed enhancement method gives a Mean Opinion Score of 3.3 as measured using the Perceptual Speech Quality Measure software.

1. INTRODUCTION

Speech enhancement methods are commonly based on speech models that have been determined from training data. In this paper, we propose a low complexity method of estimating the long-term predictor (LTP) under noisy condition. We show that adding the LTP provides significant gain of enhanced speech overall quality.

Speech models with a short-term predictor (STP) are relative simple and accurate, so they are attractive also for speech enhancement, as is reflected by the large number of publications in this area, e.g., [1], [2]. Though the estimation of the STP parameters and speech based on this model is relatively well understood (despite some open questions), it is known that usage of an autoregressive (AR) STP model only does not provide satisfactory clean-speech estimation accuracy. However, the estimation using a combined LTP and STP was addressed only in a few papers, e.g., [3]. Thus, the main goal of this paper is to compare the performance when using the STP in combination with the LTP against using only the STP. We also attempt to quantify performance of the proposed methods depending on the sample size of the STP and LTP parameters. The main difference compared to the method of [3] is that we estimate STP and LTP with *a priori* knowledge and apply some simplifications to make estimation computationaly feasible. We also use results of [4], which introduces a Kalman smoother involving a combined LTP and STP and shows that adding an LTP brings significant gain of speech quality. The difficulty when dealing with the combined LTP and STP model is the costly evaluation of the likelihood function of the parameters (the conditional probability of the noisy observation given parameters). The only method to accomplish exact likelihood evaluation with the combined LTP and STP parameters currently known is based on Kalman recursions, a method that is computationally complex since we need to handle the state error covariance matrices of size $L_{\max}+p+q$ (quadratic matrices), where $L_{\max}$ is the maximum pitch lag possible, p is the STP order, and q is the noise AR model order.

To estimate the LTP efficiently, we rely on an approximate likelihood maximization over a random codebook of LTP taps and excitation variances as well as LTP lags. The approximation applied is that the LTP parameters change once per frame (in speech coding they change once per subframe) and that the likelihood function is computed as if the frame length is infinite. Another introduced simplification is that we first estimate the STP with the assumption that the short-term residual is Gaussian and then use the resulting STP parameters to estimate the LTP. The latter simplification significantly reduces the computational complexity compared to simultaneous integration over the LTP and STP parameters space especially if dual or triple frame segments are taken into account. While estimating the LTP, we assume that the long-term residual is Gaussian distributed. It is likely that this simplification does not impair the performance significantly. Our methods have similarities to those used in speech coding. In speech coding the STP is also estimated first under assumption that the excitation is Gaussian and then the LTP is estimated from the short-term residual. The joint optimization of STP and LTP was shown to provide no advantage [5].

In this work, we assume that the noise can be described by a slowly varying AR model. To estimate the noise AR parameters we rely on the minimum statistics algorithm of Martin [6].

The main goal of the paper is to show that it is possible to devise a combined LTP and STP estimation algorithm for noisy conditions that provides a clear advantage over enhancement based on the STP only. A secondary goal is the examination of the importance of the sample size of the STP parameters used for stochastic integration and the number of frames used for MMSE estimation in the LSF domain.

2. ESTIMATION OF THE STP

We estimate the STP and LTP parameters sequentially. In this section, we describe the estimation of the STP parameters, using a method similar to that presented in [7].

The observed noisy signal frame sequence $\{\mathbf{r}^{(m)}\}$ is given by:

$$\mathbf{r}^{(m)} = \mathbf{s}^{(m)} + \mathbf{n}^{(m)}, \qquad (1)$$

where $\mathbf{s}^{(m)} = [s_{(m-1)S+1},\ldots,s_{(m-1)S+N}]^T$,

$\mathbf{n}^{(m)} = [n_{(m-1)S+1},...,n_{(m-1)S+N}]^T$ are statistically independent speech and noise random vectors corresponding to the m'th noisy speech signal frame of length N, s_t and n_t are speech and noise samples at time instant t, and S denotes the delay between the consecutive frames and is less than N.

From our asymptotic mean stationarity (AMS) assumption, we have existence of $p(\mathbf{s}^{(m-1,m,m+1)})$, $p(\mathbf{s}^{(m-1,m)})$ and $p(\mathbf{s}^{(m)})$ where $\mathbf{s}^{(m-1,m,m+1)} = [\mathbf{s}^{(m-1)}, \mathbf{s}^{(m)}, \mathbf{s}^{(m+1)}]$ is the triple frame segment, $\mathbf{s}^{(m-1,m)} = [\mathbf{s}^{(m-1)}, \mathbf{s}^{(m)}]$ is the dual frame segment. The single, dual and triple frame segments of the noisy speech are defined analogously. Let $\theta_s^{(m)}$ be the vector of the STP parameters in the *m*'th frame containing the excitation variance of the STP synthesis filter and *p* STP coefficients in the LSF form, $\theta_s^{(m-1,m)}$ be the joint parameter vector in two consecutive frames and $\theta_s^{(m-1,m,m+1)}$ be the joint parameter vector in three consecutive frames. The LSFs possess the property that all minimum phase AR polynomials result in a vector of increasingly ordered LSFs coefficients $l_1 < l_2 < ... < l_p$ [8]. It is obvious that any linear combination of the ordered LSF vectors is also ordered. This is a desirable property as the MMSE estimate is computed through linear combination of a set of the LSF vectors taken from the parameters region of support and the property assures the estimated synthesis filters to be stable. Since the parameters are obtained through a deterministic transformation from the space of speech segments they inherit the probability densities and hence we have the existence of the densities $p(\theta_s^{(m-1,m,m+1)})$, $p(\theta_s^{(m-1,m)})$ and $p(\theta_s^{(m)})$. We assume that noise parameters are known (estimated in our implementation by the *minimum statistics* method [6]) and equal θ_n. The computation of $p(\mathbf{r}^* \mid \theta_s^*, \theta_n)$ is based on the Gaussian assumption and the circulant approximation of Toeplitz matrices and other properties of the circulant matrices, see [9] and [10]. The MMSE estimation of the STP parameters is, for the single frame estimation

$$\hat{\theta}_s^{(m)} = \int_{\Omega_s} \hat{\theta}_s^{(m)} p(\theta_s^{(m)}, \theta_n \mid \mathbf{r}^{(m)}) d\theta_s^{(m)}, \tag{2}$$

for the dual frame estimation

$$\hat{\theta}_s^{(m-1,m)} = \int_{\Omega_s \times \Omega_s} \theta_s^{(m-1,m)} p(\theta_s^{(m-1,m)}, \theta_n \mid \mathbf{r}^{(m-1,m)}) d\theta_s^{(m-1,m)} \tag{3}$$

and for the triple frame estimation

$$\hat{\theta}_s^{(m-1,m,m+1)} = \int_{\Omega_s \times \Omega_s \times \Omega_s} \theta_s^{(m-1,m,m+1)} p(\theta_s^{(m-1,m,m+1)}, \theta_n \mid \mathbf{r}^{(m-1,m,m+1)}) d\theta_s^{(m-1,m,m+1)}. \tag{4}$$

To evaluate the expressions for the MSEE, we rely on the law of large numbers and on Bayes rule. We collect a data base with M entries for the single, dual and triple STP parameters distribution, using the autocorrelation method of linear prediction applied to the clean speech. The sampling is performed by randomly selecting (uniform distribution) speech segments from a large database containing over 10^7 frames. The elements of the sample are denoted as $\theta_s^*(i)$, where $*$ is either a single (m), dual $(m-1,m)$ or triple frame $(m-1,m,m+1)$ indicator. The parameter estimates are computed as:

$$\hat{\theta}_s^* = \sum_{i=1}^{M} \theta_s^*(i) p(\theta_s^*(i), \theta_n \mid \mathbf{r}^*) = \sum_{i=1}^{M} \theta_s^*(i) p(\mathbf{r}^* \mid \theta_s^*(i), \theta_n) \left(\sum_{i=1}^{M} p(\mathbf{r}^* \mid \theta_s^*(i), \theta_n) \right)^{-1}, \tag{5}$$

where the last equality follows from the application of Bayes rule. After estimating the STP parameters using these formulas, we take the estimated STP parameters for the *m*'th frame and proceed to the estimation of the LTP.

3. ESTIMATION OF THE LTP

The combined LTP and STP transfer function is

$$f(z) = \frac{\sigma_s}{1 + bz^{-L}} \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}}, \tag{6}$$

where σ_s is the excitation standard deviation of the cascade of the LTP and STP, b is the LTP tap, L is the LTP lag, $[a_p,...,a_1]$ are the STP coefficients and p is the STP order.

3.1 Preparation of the LTP parameters database

To estimate the LTP we first sample from the speech training set the LTP tap and excitation variance. We choose randomly a single frame from the training set. We then compute the STP parameters for this frame. Next, we perform analysis filtering of that frame with the zero initial conditions of the analysis filter and get the short-term residual. Extending the short-term residual by zeros at the beginning we perform, once per frame, the open-loop estimation of the LTP. The outlined procedure is repeated *K* times to collect a sample of the LTP taps and excitation standard deviations $[b(i), \sigma_s(i)]$, $i = 1,...,K$.

3.2 Estimation procedure

Let $[\alpha_1,...,\alpha_{p+1}] = [1, a_1,...,a_p]$ be the speech AR coefficients vector in the *m*'th frame obtained during STP estimation and $[\beta_1,...,\beta_{q+1}] = [1, c_1,...,c_q]$ be the noise AR coefficients vector in the *m*'th frame, which is known (in our implementation estimated by *minimum statistics* approach [6]) and σ_n be the standard deviation of the noise AR synthesis filter excitation also assumed to be known. Let us denote by

$$f_r(k) = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} r_{(m-1)S+l} w^{-(k-1)(l-1)}, \; k = 1,...,N \tag{7}$$

the DFT of the noisy speech vector in the *m*'th frame, where $w = e^{\frac{2\pi}{N} i}$,

$$f_n(k) = \frac{\sigma_n}{\sum_{l=1}^{q+1} \beta_l w^{-(l-1)(k-1)}}, \tag{8}$$

is the noise model spectrum, and

$$f_s^{(i,L)}(k) = \frac{\sigma_s(i)}{1 + b(i) w^{-(L-1)(k-1)}} \frac{1}{\sum_{l=1}^{p+1} \alpha_l w^{-(l-1)(k-1)}}, \tag{9}$$

is the speech model spectrum. Using this notation, the estimation proceeds as follows:

for $L = L_{\min},...,L_{\max}$ (LTP lags)
 for $i = 1,...,K$ (pairs of LTP taps and excitation variances)

$$D_{L,i} = -\frac{1}{2}\sum_{k=1}^{N}\log\left(\left|f_s^{(i,L)}(k)\right|^2 + |f_n(k)|^2\right) - ...$$

$$+\frac{1}{2}\sum_{k=1}^{N}\frac{|f_r(k)|^2}{\left|f_s^{(i,L)}(k)\right|^2 + |f_n(k)|^2}$$

 end
end

The algorithm searches for the maximum of the likelihood function. The obtained maximum is global over the discrete set. The method assumes the excitation of the speech model from (6) to be i.i.d. Gaussian and that the circulant approximation to the Toeplitz matrices is sufficiently accurate. The initial condition for each frame is set to zero. In practice, the method selects the L, $b(i)$ and $\sigma_s(i)$ that maximise $D_{L,i}$.

4. SPEECH ENHANCEMENT USING WIENER FILTER OR KALMAN SMOOTHER

We performed experiments with a Wiener filter and a Kalman smoother. First we review the Wiener filtering method. Denote the estimated speech model spectrum, as given in the previous section, as $\hat{f}_s(k)$. The periodogram of the estimated speech is:

$$\left|\hat{S}(k)\right|^2 = \frac{\left|\hat{f}_s(k)\right|^2}{\left|\hat{f}_s(k)\right|^2 + |f_n(k)|^2}|f_r(k)|^2. \qquad (10)$$

The spectrum of the estimated speech frame is synthesised using the phase spectrum of the noisy frame, that is:

$$\hat{S}(k) = \left|\hat{S}(k)\right|\exp(i\angle f_r(k)). \qquad (11)$$

The estimated speech *m*'th frame is given by the inverse DFT:

$$\hat{s}((m-1)S + l) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\hat{S}(k)w^{(l-1)(k-1)}. \qquad (12)$$

Note that the segments of the estimated speech overlap on the frame boundaries on the time segment equal $O = N - S$. Hence, the estimated speech segments are weighted using a half of $\sin^2$ cycle at the beginning and a half of $\cos^2$ cycle at the end before the overlap-add.

The Kalman fixed lag smoother computes clean speech estimates according to the following prescription:

$$\hat{s}_{t-L_{\max}-p} = E\left[s_{t-L_{\max}-p} \mid r_t, r_{t-1},...\right]. \qquad (13)$$

The parameters from the *m*'th frame are fed into the Kalman smoother for $t = (m-1)S + \frac{O}{2} + 1,...,(m-1)S - \frac{O}{2} + N$. Details about how to perform the Kalman smoothing with the combined LTP and STP are contained in [4].

5. EXPERIMENTAL STUDY

The first set of experiments was aimed at finding the dependency between length of the segments (single, dual or triple, cf. section 2.), sample size (parameter *M,* cf. section 2.) and the accuracy of the AR parameters estimation as measured using the root mean spectral distortion (SD). The experiments were done for a benchmark file of speech mixed with vehicle noise (Volvo noise taken from the NOISEX-92 database) at -5dB. We chose car noise for the experiments since the car application generally allows for higher computational complexity.

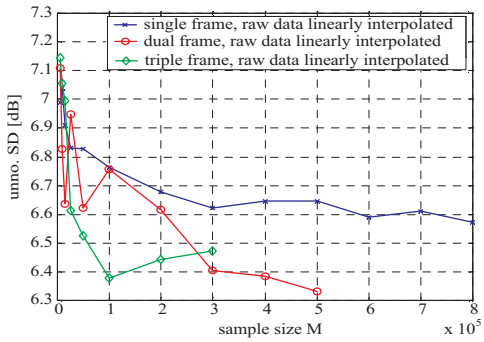

Figure 1. Unnormalised spectral distortion (comparing spectral shapes and the excitation variances) depending on the length of the segments and the sample size used to average the MMSE estimates.

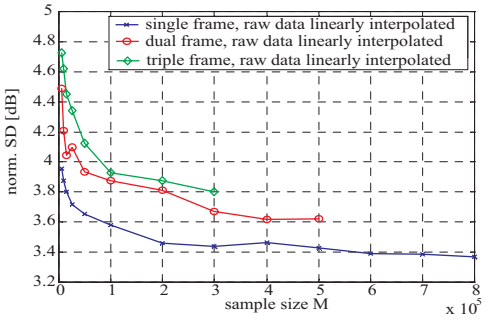

Figure 2. Normalised spectral distortion (comparing only the spectral shapes) depending on the length of the segments and the sample size used to average the MMSE estimates.

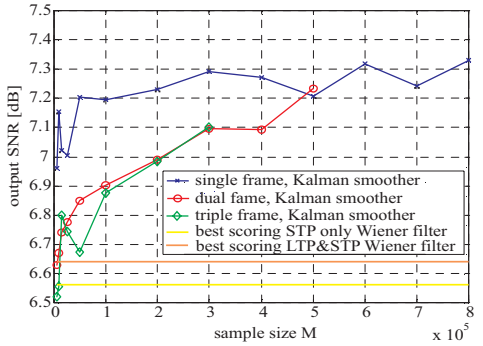

Figure 3. SNR of the enhanced sequence, obtained at fixed *K*=500, depending on the length of the segments and the sample size used to average the MMSE estimates. The Wiener and Kalman filter uses the STP parameters computed according to the recipe from the Section 2.

We used a frame length *N* = 256 and a step size of *S* =236. The test sequence was 12.5 [s] containing speech material outside the training set. The sampling frequency was 8 kHz. The training set was the TIMIT database. To compute the SD we used only frames with a mean power larger than 15dB below the mean power of the entire test sequence (the test sequence had pauses between sentences maximally of the length of 0.2 [s]). We

excluded the silent frames since these contribute much to the SD but are not significant to perception. Due to limitations on computing time we did not test the dual frame estimation for sample size larger than M = 500000 and the triple frame estimation for sample size larger than M = 300000. We did not, for the same reason, average the results over several collections of samples.

The SNR and PSQM results were obtained for a fixed size, K = 500 (we recall that K is the number of pairs of the LTP taps and long-term residual variances over which the likelihood was maximized), of the LTP parameters sample. Before computing the PSQM scores, we filtered the test sequence and the enhanced files with a high pass filter with a cut-off frequency of 110 Hz. The yellow line in Figs. 3. and 4. shows the result of the best performing Wiener filter with the STP speech model estimated as in the Section 2. The remaining lines are obtained with the combined LTP and STP speech model.

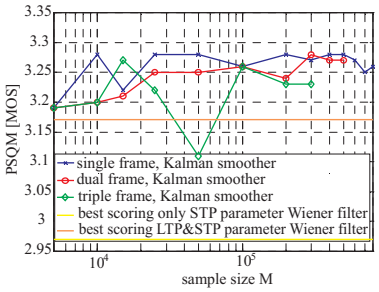

Figure 4. PSQM measurements, obtained at fixed K=500, depending on the length of the segments and the sample size used to average the MMSE estimates. The Wiener and Kalman filter uses the STP parameters computed according to the recipe from the Section 2.

To determine the performance of the proposed algorithm as a function of the size of the LTP parameters sample, we ran experiments with fixed M = 50000. The results are shown in the Fig. 5.

We also measured the SNR and PSQM performance for a spectral subtraction algorithm and a standardized EVRC noise suppression system [11]. The results are summarized in Table 1. It is seen that our method provides 0.7 improvement on the MOS scale over the EVRC noise suppression.

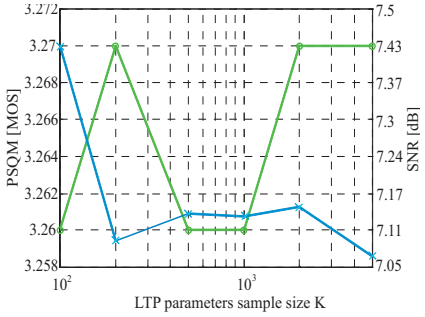

Figure 5. PSQM (left scale, green line, circles) and SNR (right scale, blue line, crosses) measurements depending on size of the LTP parameters sample K.

Table 1. Performance measurements for some other speech enhancement systems. All results was obtained from the same -5dB, speech mixed with vehicle noise, input sequence.

	SNR_{out} [dB]	$PSQM_{out}$ [MOS]
Spectral subtraction	6.1	1.83
Ramabadran, see [11]	6.2	2.58

6. CONCLUSIONS

We observed a significant gain in the enhanced speech quality when using a combined LTP and STP model over using only an STP model. In terms of PSQM measurements this improvement is 0.3 on the MOS scale and informal listening tests confirm this result. Conditioning of the STP estimation using neighboring frames is advantageous only for unnormalized SD measurements. This result implies that the excitation variances of the triple-frame STP model were obtained with higher accuracy than those of the single-frame estimation scheme. The spectral shapes were estimated with higher accuracy with the single frame STP model. One of the main results of our investigation is that adding context to the STP estimation does not provide better estimation accuracy (at least for the sample size used in our experiments).

Increasing the sample size M over M = 100000 does not result in a significant gain of quality as measured using PSQM. An increase of the LTP parameters sample size K gave little or no improvement in PSQM and SNR. A related issue is that our method, despite good PSQM scores, currently introduces some impulsive distortions. These distortions occur locally in the enhanced files and do not affect the PSQM measurements significantly but are annoying. The distortions result from misestimated LTP parameters. These results are consistent with the block size of the LTP being a limiting factor. We plan to eliminate the problems by introducing improved LTP modelling.

7. REFERENCES

[1] J.S. Lim, A.V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. ASSP-26*, 1978, pp. 197-209.

[2] M. Gabrea, "Robust adaptive Kalman filtering-based speech enhancement algorithm," *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 1, 2004, pp. 301-304.

[3] M. Kuropatwinski, D. Leckschat, K. Kroschel, A. Czyzewski, and C. Hales, "Speech Enhancement for Linear Predictive Analysis by Synthesis Coders," *Proc. Eurospeech*, 1999, pp. 2383-2386.

[4] M. Kuropatwinski, D. Leckschat, K. Kroschel and A. Czyzewski, "Integration of Speech Enhancement and Coding Techniques," *Proc. IEEE Speech Coding Workshop*, 1999, pp. 160-162.

[5] R.V. Ramachandran, P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. ASSP*, vol 37, 1989, pp. 467-476.

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

[7] M. Kuropatwinski and W.B. Kleijn, "MMSE Estimation of the Short Term Predictor Parameters under Noisy Conditions," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2003, pp. 1.96-1.99.

[8] F. Soong and B. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, San Diego, 1984, pp.1.10.1-1.10.4.

[9] P.J. Davis, *Circulant Matrices.* Wiley, New York, 1979.

[10] S. Srinivasan, J. Samuelsson and W.B. Kleijn, "Estimation of short-term predictor parameters for coding and enhancement of noisy speech," *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 1, 2004, pp. 705-708.

[11] T. Ramabadran, J.P. Ashley, and M.J. McLaughlin, "Background Noise Suppression for Speech Enhancement and Coding," *Proc. IEEE Speech Coding Workshop,* 1997, pp. 43-44.