ROBUST SPECTRUM QUANTIZATION FOR LP PARAMETER ENHANCEMENT

Volodya Grancharov, Sriram Srinivasan, Jonas Samuelsson and W. Bastiaan Kleijn

KTH (Royal Institute of Technology)
Department of Signals, Sensors and Systems
PSfrag replacements 10044 Stockholm, Sweden
{volodya.grancharov, sriram.srinivasan, jonas.samuelsson, bastiaan.kleijn}@s3.kth.se

ABSTRACT

In this paper, we investigate the denoising properties of robust vector quantization of the speech spectrum parameters in combination with a Kalman filter. The underlying assumption is that the high-energy speech regions can be used to reconstruct the low-energy regions destroyed by noise. This can be achieved through vector quantization with a properly weighted distortion measure. The performance of the proposed system, Kalman filtering with prior vector quantization, is compared with existing schemes for parameter estimation used in Kalman filtering. The results indicate significant improvement over the reference systems in both objective and subjective tests.

1. INTRODUCTION

Let the speech $s(n)$ be observed in the presence of additive colored noise process $v(n)$:

$$y(n) = s(n) + v(n). \quad (1)$$

Under practical conditions, the performance of the speech communication system can be heavily affected by acoustic background noise. The noise suppression may be of different types [1], [2], but is usually a pre-processor to the speech codecs. Such a system operates without exploiting a-priori speech information present in the codec, such as the codebook of linear prediction (LP) coefficients, as illustrated in Figure 1.

Accurate estimation of speech and noise parameters has a crucial role for the noise suppression system. This is especially valid for the single-channel application, where the estimation of these parameters is considered as a more complex problem than the noise suppression itself. The noise statistics can be estimated using a voice activity detector or soft decision methods such as quantile based [3] or minimum statistics based noise estimation [4]. For model-based noise suppression systems such as the Kalman filter, the speech parameters should also be estimated. The existing methods are based on spectral subtraction schemes, iterations [5] or Expectation-Maximization algorithms [6].

Conventional noise suppression systems do not exploit the fact that the speech energy is non-uniformly distributed over the frequencies. Therefore, even though high-frequency regions may be completely destroyed by the noise, in the low-frequency regions, the signal-to-noise ratio (SNR) level may be relatively high, as illustrated in Figure 2. Possessing a-priori speech information in the form of a speech codebook makes it possible to reconstruct low-energy regions, based on the information from the relatively well-preserved high-energy regions. In the next section, we propose a scheme to incorporate the "denoising" properties of vector quantization in a conventional noise suppression system. Without loss of generality, we shall assume that the vector quantization operates on the polynomial LP coefficients $\mathbf{a}_s$.

This work was partially supported by the European Commission under the ANITA project (IST-2001-34327).

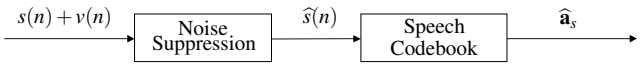

Figure 1: The standard configuration of a noise suppression system followed by LP quantization.

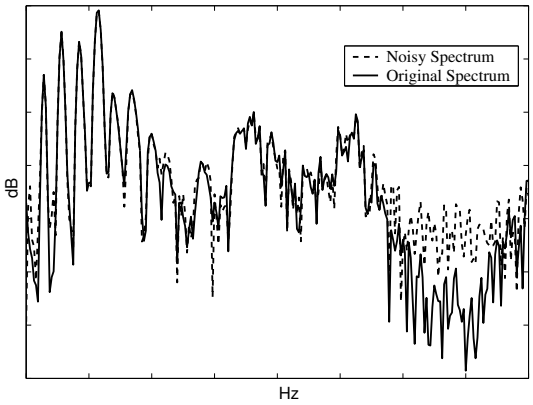

Figure 2: A plot of the original and noisy spectral envelopes demonstrating that the SNR varies strongly with frequency.

2. SPEECH ENHANCEMENT WITH PRIOR SPECTRUM QUANTIZATION

We study the effectiveness of robust vector quantization as an enhancement technique for speech LP parameters for use in Kalman filtering. In the proposed algorithm, the LP spectrum of the noisy signal is first quantized with a robust-to-noise distortion measure. The quantized and "enhanced" LP coefficients are used in the Kalman filter as model parameters, as illustrated in Figure 3. The ideas of joint waveform enhancement and spectrum quantization can already be found in the literature, for example [7]. The authors use the conventional Itakura-Saito distortion measure to iteratively design a Wiener filter using a speech codebook. In contrast to our work, this approach does not exploit unequal SNR conditions.

lacements

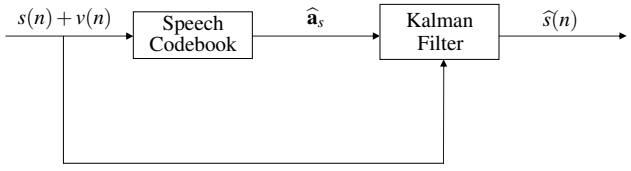

Figure 3: Kalman filter with prior quantizer for enhancement of speech model parameters.

2.1 Kalman filtering approach to speech enhancement

In this section we will closely follow the notation used in [5]. To apply the Kalman filter, we model the speech and noise as autoregressive processes of model order p and q respectively:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + w(n) \tag{2}$$

and

$$v(n) = \sum_{k=1}^{q} b_k v(n-k) + u(n), \tag{3}$$

where $w(n)$ and $u(n)$ are white noise sequences. The speech and the noise model orders are set to $p = 10$ and $q = 10$. The system of equations (1-3) can be represented in an extended state-space form:

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{F}\,\mathbf{x}(n-1) + \mathbf{G}\,\mathbf{z}(n) \\ y(n) &= \mathbf{H}^T\,\mathbf{x}(n), \end{aligned} \tag{4}$$

$\mathbf{x}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-p+1)\ v(n)\ v(n-1)\ \ldots\ v(n-q+1)]^T$ is the $(p+q)$ dimensional state vector and $\mathbf{z}(n) = [w(n)\ u(n)]^T$. The explicit expressions for $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{H}$ are given below

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_s & \mathbf{0}_{p,q} \\ \mathbf{0}_{p,q} & \mathbf{F}_v \end{pmatrix}$$

$$\mathbf{G} = \begin{pmatrix} 1\,0\,\cdots\,0 & 0\,0\,\cdots\,0 \\ \underbrace{0\,0\,\cdots\,0}_{p} & \underbrace{1\,0\,\cdots\,0}_{q} \end{pmatrix}^T$$

$$\mathbf{H} = \begin{pmatrix} \underbrace{1\,0\,\cdots\,0}_{p} & \underbrace{1\,0\,\cdots\,0}_{q} \end{pmatrix}^T,$$

and the transition matrices for speech and noise are given by:

$$\mathbf{F}_s = \begin{pmatrix} a_1 & a_2 & \cdots & a_{p-1} & a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$\mathbf{F}_v = \begin{pmatrix} b_1 & b_2 & \cdots & b_{q-1} & b_q \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Using the state-space representation (4), the Kalman filter estimate becomes [5], [8]

$$\begin{aligned} \hat{\mathbf{x}}(n) &= \mathbf{F}\hat{\mathbf{x}}(n-1) + \mathbf{k}(n)[y(n) - \mathbf{H}^T\mathbf{F}\hat{\mathbf{x}}(n-1)] \\ \mathbf{k}(n) &= \mathbf{P}(n|n-1)\mathbf{H}[\mathbf{H}^T\mathbf{P}(n|n-1)\mathbf{H}]^{-1} \\ \mathbf{P}(n|n-1) &= \mathbf{F}\mathbf{P}(n-1)\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T \\ \mathbf{P}(n) &= [\mathbf{I} - \mathbf{k}(n)\mathbf{H}^T]\mathbf{P}(n|n-1), \end{aligned}$$

where $\mathbf{k}(n)$ is the Kalman gain and $\hat{\mathbf{x}}(n)$ is the estimate of the state $\mathbf{x}(n)$. $\mathbf{P}(n|n-1) = E\{(\mathbf{x}(n) - \hat{\mathbf{x}}(n))(\mathbf{x}(n) - \hat{\mathbf{x}}(n))^T\}$ is the prediction-error covariance matrix and $\mathbf{P}(n) = E\{(\mathbf{x}(n) - \mathbf{F}\hat{\mathbf{x}}(n-1))(\mathbf{x}(n) - \mathbf{F}\hat{\mathbf{x}}(n-1))^T\}$ is the filtering-error covariance matrix, the noise covariance is given by $\mathbf{Q} = E\{(\mathbf{z}(n))(\mathbf{z}(n))^T\}$. The speech sample estimate can be obtained by $\hat{s}(n) = [1\ 0\ \cdots\ 0]_{p+q}\hat{\mathbf{x}}(n)$. On a frame-by-frame basis the values of $\mathbf{F}_s$, $\mathbf{F}_v$ and $\mathbf{Q}$ are updated and the Kalman filter is re-initialized. In our implementation the minimum statistics method [4] was used to estimate the noise statistics. The speech parameters were estimated by the novel algorithm presented in the next section.

2.2 Robust quantization of the LP parameters

To achieve the "denoising" effect from the quantization operation, in Figure 3, the codebook should be pre-trained with clean speech. Quantization however is performed on a noisy input signal. The LPC analysis does not explicitly account for noise, and the different noise conditions lead to a mismatch in the LPC spectrum of the test and reference model. Quantizing with conventional measures, such as spectral distortion (SD) typically fails in low SNR conditions. From Figure 2, we can conclude that the mismatch in the valleys will make a conventional distortion measure large and decrease the probability of selection of the coded entry corresponding to the best clean-speech match. To exploit the relatively higher noise immunity exhibited by the spectral peaks, we need to find a distortion measure that weights the error in the peaks more than the error in the valleys. One such measure that can be found already in the literature is the weighted Itakura distortion [9].

2.2.1 Weighted Itakura distortion

The gain-normalized weighted Itakura distortion measure, [9] is defined as:

$$d_{WI} = \log \int_{-\pi}^{\pi} F(\omega) \frac{|B(\omega)|^2}{|A(\omega)|^2} \frac{d\omega}{2\pi}, \tag{5}$$

where the test LP spectrum $A(\omega)$ is calculated as the Fourier transform of the linear prediction polynomial $(1, a_1, \ldots, a_p)$, calculated over a short speech segment of the signal incoming to the system. Similarly, the reference LP spectrum $B(\omega)$ is calculated as the Fourier transform of the linear prediction polynomial $(1, b_1, \ldots, b_p)$, stored in the codebook. The weighting function $F(\omega)$ is of the form:

$$F(\omega) = \frac{1}{|A_\alpha(\omega)|^2} = \frac{1}{|1 + \sum_{k=1}^{p} \alpha^k a_k e^{-j\omega k}|^2}, \tag{6}$$

where $\alpha \in [0,1]$ is the bandwidth-broadening factor. Any nonzero value of α produces a weight that is larger at the peaks and smaller in the valleys. In noise-free conditions,

$\alpha = 0$ and the errors from the valleys and the peaks are equally weighted. An example for a weighting function is given in Figure 4. In the extreme case when $\alpha = 1$, the weighting function will be the test spectrum itself. The emphasis factor α can be estimated using grid search for a given noise type and level. The search can minimize an appropriate distortion measure, or the error rate of a speech recognition system, as in [9].

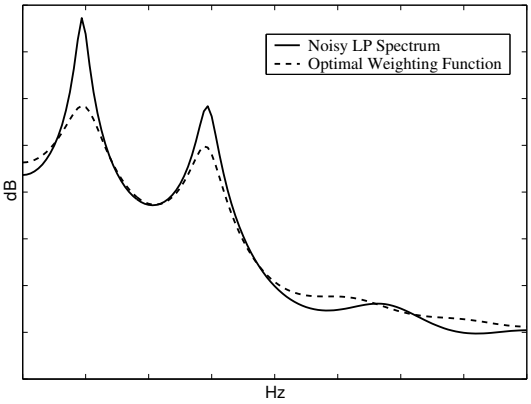

Figure 4: Noise immunity of the spectral peaks exploited by weighting function $F(\omega)$ with $\alpha = 0.8$.

2.2.2 *Extension for the colored noise*

The discussion in [9] is limited to the case of white Gaussian noise and the optimal α is dependent on the SNR values only. Real noise sources often have a colored spectrum and for good performance, the emphasis factor α should account for that. For example, the auto-regressive speech envelope can be destroyed by white noise at 5 dB SNR, but is almost not affected by car noise at the same SNR level. The reason is that most of the noise energy in the second case is concentrated in the non-audible regions.

This motivates the introduction of one more parameter in the noise model: the noise spectrum tilt. Such a noise model is more flexible, but still simple enough to use the off-line algorithm to find the optimal emphasis factor. The mapping is made from the parameter pair {SNR, noise spectrum tilt} to α, instead of mapping the single parameter {SNR} to α. The noise spectrum tilt is calculated as the coefficient in the first order linear prediction polynomial, calculated from the estimate of the noise spectrum.

The pre-training for obtaining the optimal emphasis factor was done similarly to [9], except that the noise tilt was varied at each SNR level. This was done by creating an artificial white Gaussian noise with a variance that leads to a given overall SNR. Before addition to the clean speech, the generated white noise was pre-filtered to obtain the desired spectrum tilt. Ten sentences from the TIMIT database [10] were quantized with the weighted-Itakura distortion measure. A grid search was performed over the values of α and the optimal emphasis factor was chosen as the value minimizing the overall SD between clean and quantized LP spectra. The step size was chosen to be 5 dB for the SNR and 0.1 for α. A 10-bit speech codebook was trained using the generalized Lloyd algorithm with 10 minutes of speech from TIMIT database using the Itakura-Saito distortion measure [11]. The LP coefficients were extracted from the signal every 30 ms with 50% overlap. The SD was chosen as a measure that evaluates the closeness between the clean speech auto-regressive envelope $A(\omega)$ and the auto-regressive envelope of the processed signal $\widehat{A}(\omega)$ [12]. For the n^{th} frame the instantaneous SD is given by:

$$SD_n^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(10\log_{10}|A(\omega)|^2 - 10\log_{10}|\widehat{A}(\omega)|^2\right)^2 d\omega.$$

For the evaluations the instantaneous SD values are averaged over all speech frames by:

$$\overline{SD} = \frac{1}{M}\sqrt{\sum_{n=1}^{M} SD_n^2},$$

where M is the total number of frames. This procedure results in a lookup table for SNR, tilt and the corresponding α. In the training phase the system was presented with the "ideal" noise parameters. The resulting lookup table is presented in Table 1. The complete system consists of a noise

	Tilt			
SNR(dB)	0.0	0.3	0.6	0.9
5	0.9	0.9	0.6	0.1
10	0.8	0.7	0.3	0.1
15	0.7	0.5	0.3	0.1
20	0.4	0.2	0.2	0.0

Table 1: Optimal emphasis factor for a given SNR and noise spectrum tilt.

estimation algorithm that provides the SNR and the noise spectral tilt to the lookup procedure, and a quantization with the weighted Itakura distortion measure, see Figure 5.

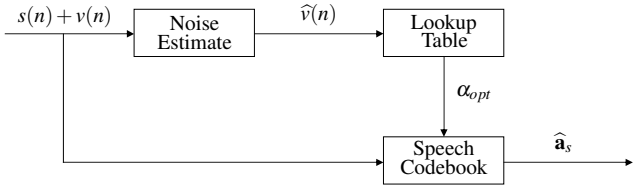

Figure 5: Robust vector quantization with frequency-weighted Itakura distortion measure.

3. PERFORMANCE

In this section we evaluate the performance of the vector quantization with the weighted-Itakura distortion measure. Since our quantization method enhances only the speech spectral envelope, the evaluation was performed in terms of SD values. Later we integrate the vector quantization in the Kalman filter speech enhancement system and measure the performance against reference systems in terms of preference listening tests.

3.1 Evaluation of LP parameter enhancement through robust vector quantization

We evaluated the advantages of using the weighted Itakura measure in comparison to quantization with a conventional

distortion measure, such as Itakura-Saito. The same settings were used as for the pre-training phase, except that the minimum statistics algorithm was used for noise parameter estimations [4]. The clean speech was contaminated with white Gaussian and rain noise, each at 5 dB and 10 dB SNR level. The rain noise spectrum exhibits a strong peak and therefore it is a difficult source for the chosen noise model {SNR, noise spectrum tilt}. The noisy input was quantized with either the weighted Itakura or Itakura-Saito measure and the result was evaluated in terms of SD. The results, averaged over ten speech sentences, are presented in Table 2. To confirm the soundness of these improvements, in the next subsection we used the enhanced LP parameters as speech model parameters for waveform enhancement through Kalman filtering.

Noise Type	SNR Level	Reduction in SD(dB)
White	5 dB	0.8
	10 dB	0.7
Rain	5 dB	0.4
	10 dB	0.4

Table 2: Performance in SD for the conventional and weighted Itakura distortion measure.

3.2 Comparison with other enhancement methods

We performed simulations to evaluate the effectiveness of the proposed system in comparison with some of the existing algorithms for parameter enhancement for Kalman filtering. The Kalman filter used in the tests was implemented according to section 2.1. As a first reference system for LP parameter enhancement, we implemented the iterative algorithm proposed in [5]. In the first iteration, the LP coefficients were obtained from the noisy signal. The output of the Kalman filter becomes the input for the second iteration. Only three iterations were used, since the quality deteriorates for further iterations, as observed in [5]. In the second reference system, the estimated noise power spectrum was subtracted from the noisy power spectrum (spectral subtraction) and the LP coefficients were calculated from the resulting estimate of the clean spectrum.

Eight listeners not familiar with the systems participated in the tests. Listeners were provided with two utterances at a time and were asked to choose the one with higher quality. The test material consisted of five male and five female speakers arbitrary chosen from the TIMIT database and two different noise sources. The results for the preference listening tests, presented in Table 3, indicate a clear advantage of the proposed algorithm over the reference systems.

Noise Type	System type	Preference
Rain 10 dB	Iterative KF	78.7%
	Spectral subtraction + KF	83.8%
White 10 dB	Iterative KF	77.5%
	Spectral subtraction + KF	82.5%

Table 3: Listeners' preference for the system weighted Itakura + Kalman filter (KF) against competitive parameter enhancement schemes.

4. CONCLUSIONS

Vector quantization with a robust-to-noise distortion measure is a simple and efficient system for parameter enhancement. Joint use of a model-based noise suppression system and vector quantization that exploits the a-priori speech information available in the speech codebook is a promising approach. The main advantage of the proposed algorithm is that the codebook poses a constraint on the choice of model parameters. The tests confirm that this leads to improved performance over conventional parameter enhancement schemes used for Kalman filtering.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109 –1121, Dec. 1984.

[2] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 12, pp. 177– 180, Apr. 1987.

[3] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 3, pp. 1875–1878, June 2000.

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

[5] J. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, pp. 1732–1742, Aug. 1991.

[6] S. Gannot, D. Burshtein, and E.Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[7] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 4, no. 5, pp. 383–389, Sept. 1996.

[8] T. Kailath, A. Sayed, and B. Hassiby, *Linear Estimation*. New Jersey: Prentice Hall, 2000.

[9] F. Soong and M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 41–48, Jan. 1988.

[10] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1-1.1*, 1990.

[11] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[12] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Amsterdam: Elsevier Science Publishers, 1995.