

Speech Enhancement Using A-Priori Information

Sriram Srinivasan, Jonas Samuelsson and W. Bastiaan Kleijn

Speech Processing Group, Dept. of Signals, Sensors and Systems
Royal Institute of Technology (KTH), Stockholm, Sweden

{sriram.srinivasan, jonas.samuelsson, bastiaan.kleijn}@s3.kth.se

Abstract

In this paper, we present a speech enhancement technique that uses a-priori information about both speech and noise. The a-priori information consists of speech and noise spectral shapes stored in trained codebooks. The excitation variances of speech and noise are determined through the optimization of a criterion that finds the best fit between the noisy observation and the model represented by the two codebooks. The optimal spectral shapes and variances are used in a Wiener filter to obtain an estimate of clean speech. The method uses both a-priori and estimated noise information to perform well in stationary as well as non-stationary noise environments. The high computational complexity resulting from a full search of joint speech and noise codebooks is avoided through an iterative optimization procedure. Experiments indicate that the method significantly outperforms conventional enhancement techniques, especially for non-stationary noise.

1. Introduction

Enhancement of speech corrupted by background noise is a topic of long standing interest as it has applications in a wide range of areas. Numerous noise suppression techniques such as Wiener filtering, subtractive type methods and subspace based methods have been developed. However, a common feature of most single channel speech enhancement systems is that they require some form of noise estimation to obtain information about noise statistics. These estimation techniques include voice activity detection, estimation from initial silence segments, and more recently methods based on quantiles [1] and minimum statistics [2]. While the recent noise estimation techniques are designed to perform well even in non-stationary noise environments, performance still degrades with increasing non-stationarity.

One way to overcome this problem is by using a-priori knowledge about speech and noise. A solution based on this principle was presented in [3] where a-priori information is stored in trained codebooks. The method uses two codebooks, one each for speech and noise auto-regressive (AR) spectral envelopes. For a given noisy frame of speech, for each speech and noise entry from the joint codebook, the excitation variances and a likelihood score are computed. The score captures the likelihood that the observed noisy frame is generated by a given pair of speech and noise spectral shapes, together with their variances. The codebook entries and the related variances that globally maximize the likelihood score can then be used in an enhancement technique such as Kalman or Wiener filtering. A schematic diagram of this method is shown in figure 1.

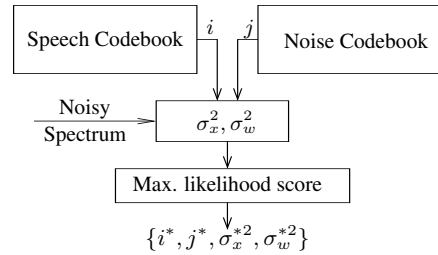


Figure 1: Estimation of excitation variances and spectral shapes: i^* , j^* are the indices of the selected entries from the speech and noise codebooks and σ_x^{*2} , σ_w^{*2} are the corresponding excitation variances.

While the method of [3] is promising, the high computational complexity incurred by the global search of the joint codebook limits the size of the speech and noise codebooks. Smaller codebooks provide poorer representation of the signals resulting in inaccurate estimation. In [3], a 10-bit speech codebook (AR model order 8) and a 4-bit noise codebook (AR model order 6) were used. In this paper, we present an iterative scheme that eliminates the need for a complete search through a joint codebook. The method uses both estimated noise information and a-priori information to reduce complexity and achieve better performance. We also present new optimization criteria and estimation of the excitation variances.

Other methods that use a-priori information include a codebook based Wiener filter approach [4] and hidden Markov model (HMM) based methods [5]. In [4], a speech codebook is used to obtain an updated estimate of the clean speech LPC vector from the current estimate at each iteration. However, the method depends on noise estimation to obtain the initial estimate of clean speech resulting in poor performance for non-stationary noise. The method proposed in this paper is different in that a-priori information about both speech and noise is employed to actively estimate the respective excitation variances and spectral shapes. In HMM based techniques, the clean signal is modelled using Gaussian AR HMMs and noise is modelled as a Gaussian process. The minimum mean-squared error (MMSE) estimator of clean speech given the noisy speech is obtained as a weighted sum of MMSE estimators corresponding to each state of the HMM for the clean signal. The method proposed in this paper and the method of [3] both differ from the HMM-based methods in that they employ trained codebooks of AR spectral shapes instead of HMMs. Many existing speech coders are linear predictive analysis by synthesis (LPAS) coders that contain codebooks of AR parameters obtained through linear prediction. The output of the proposed enhancement system can directly be used in LPAS speech coders.

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

2. Parameter estimation

Consider an additive noise model where speech and noise are independent:

$$y(n) = x(n) + w(n), \quad (1)$$

where $y(n)$, $x(n)$ and $w(n)$ represent the noisy speech, clean speech and noise respectively. For each frame, the noisy spectrum can be modelled by a combination of speech and noise AR spectral shapes from the respective codebooks, together with their excitation variances. Given the spectral shapes and excitation variances, the modelled noisy spectrum can be written as

$$\hat{P}_y(\omega) = \frac{\sigma_x^2}{|a_x(\omega)|^2} + \frac{\sigma_w^2}{|a_w(\omega)|^2}, \quad (2)$$

where σ_x^2 and σ_w^2 are the excitation variances of clean speech and noise respectively, and

$$a_x(\omega) = \sum_{k=0}^p a_{x_k} e^{-j\omega k}, \quad a_w(\omega) = \sum_{k=0}^q a_{w_k} e^{-j\omega k}, \quad (3)$$

where $\theta_x = (a_{x_0}, \dots, a_{x_p})$, $\theta_w = (a_{w_0}, \dots, a_{w_q})$ are the AR coefficients of clean speech and noise with p, q being the respective AR-model orders. The parameters to be estimated are $\{\sigma_x^2, \sigma_w^2, \theta_x, \theta_w\}$.

The above parameter estimation problem can be solved by finding the best spectral fit between the observed and the modelled noisy spectrum, with respect to a particular distortion measure. In general this is a difficult problem, but can be solved by restricting the search space using a-priori information in the form of trained codebooks of speech and noise spectral shapes. For each combination of θ_x, θ_w from the speech and noise codebooks, the excitation variances can be obtained by minimizing $d(P_y(\omega), \hat{P}_y(\omega))$, where d is the chosen distortion measure and $P_y(\omega)$ is the observed noisy spectrum. The parameter set resulting in a global minimum of $d(P_y(\omega), \hat{P}_y(\omega))$, for all codebook combinations will be the optimal solution to the estimation problem. The conditions for this technique to result in a unique solution are described in [3].

2.1. Optimization of log-spectral distortion

The log-spectral distortion between the observed noisy spectrum and the noisy spectrum obtained from the model is given by

$$d_{LS} = \frac{1}{2\pi} \int \left| \ln \left(\frac{\sigma_x^2}{|a_x(\omega)|^2} + \frac{\sigma_w^2}{|a_w(\omega)|^2} \right) - \ln(P_y(\omega)) \right|^2 d\omega. \quad (4)$$

Given $a_x(\omega)$ and $a_w(\omega)$, the corresponding optimal excitation variances can be determined by differentiating (4) with respect to σ_x^2 and σ_w^2 , setting the result to zero and solving the resulting set of simultaneous equations. First we simplify (4) to ensure that the resulting equations are linear :

$$\begin{aligned} d_{LS} &= \frac{1}{2\pi} \int \left| \ln \left(\frac{\sigma_x^2}{|a_x(\omega)|^2} + \frac{\sigma_w^2}{|a_w(\omega)|^2} \right) \right|^2 d\omega \\ &\approx \frac{1}{2\pi} \int \left| \frac{\sigma_x^2}{|a_x(\omega)|^2} + \frac{\sigma_w^2}{|a_w(\omega)|^2} - P_y(\omega) \right|^2 d\omega, \end{aligned} \quad (5)$$

where we used the approximation $\ln(1+x) \approx x$, for small x , i.e., small modelling errors. Partial differentiation with respect

to σ_x^2 and σ_w^2 yields

$$\begin{aligned} \int \frac{\sigma_x^2 |a_w|^2 + \sigma_w^2 |a_x|^2 - P_y |a_x|^2 |a_w|^2}{P_y |a_x|^2 |a_w|^2} \left(\frac{1}{P_y |a_x|^2} \right) d\omega &= 0, \\ \int \frac{\sigma_x^2 |a_w|^2 + \sigma_w^2 |a_x|^2 - P_y |a_x|^2 |a_w|^2}{P_y |a_x|^2 |a_w|^2} \left(\frac{1}{P_y |a_w|^2} \right) d\omega &= 0, \end{aligned}$$

where the dependency on ω has not been shown to facilitate notation. The resulting solution can be written as:

$$\begin{aligned} \left[\begin{array}{c} \left\| \frac{1}{P_y^2(\omega) |a_x(\omega)|^4} \right\| \left\| \frac{1}{P_y^2(\omega) |a_x(\omega)|^2 |a_w(\omega)|^2} \right\| \\ \left\| \frac{1}{P_y^2(\omega) |a_x(\omega)|^2 |a_w(\omega)|^2} \right\| \left\| \frac{1}{P_y^2(\omega) |a_w(\omega)|^4} \right\| \end{array} \right] \left[\begin{array}{c} \sigma_x^2 \\ \sigma_w^2 \end{array} \right] \\ = \left[\begin{array}{c} \left\| \frac{1}{P_y(\omega) |a_x(\omega)|^2} \right\| \\ \left\| \frac{1}{P_y(\omega) |a_w(\omega)|^2} \right\| \end{array} \right], \end{aligned} \quad (6)$$

where $\|f(\omega)\| = \int |f(\omega)| d\omega$.

The estimation procedure can be summarized as follows. For each pair of speech and noise spectral shapes from the respective codebooks, the excitation variances are calculated according to (6) and the distortion (4) is evaluated. Negative variances arising due to model errors are set to zero. The speech and noise spectra globally minimizing the distortion measure are determined. These spectra with the corresponding excitation variances represent the combination that provide the best spectral fit to the observed noisy spectrum with respect to the measure (4).

2.2. Optimization of Itakura-Saito distortion

The Itakura-Saito distortion measure between two spectral densities P and \hat{P} is defined as [6]

$$d_{IS}(P, \hat{P}) = \frac{1}{2\pi} \int (P/\hat{P}) - \ln(P/\hat{P}) - 1 d\omega. \quad (7)$$

For small distortion, using a series expansion for $\ln(x)$ up to second order terms, it has been shown that [6]

$$d_{IS}(P, \hat{P}) \approx \frac{1}{2} d_{LS}(P, \hat{P}) \quad (8)$$

Thus, for small distortions, equation set (6) gives the optimal excitation variances for the Itakura-Saito measure as well.

2.3. Optimization of the log-likelihood criterion

Under Gaussianity assumptions, the log-likelihood function to be maximized can be shown to be [3]:

$$\begin{aligned} d_{LL} &= - \int \ln \left(\frac{\sigma_x^2 |a_w(\omega)|^2 + \sigma_w^2 |a_x(\omega)|^2}{|a_x(\omega)|^2 |a_w(\omega)|^2} \right) \\ &\quad + P_y(\omega) \left(\frac{|a_x(\omega)|^2 |a_w(\omega)|^2}{\sigma_x^2 |a_w(\omega)|^2 + \sigma_w^2 |a_x(\omega)|^2} \right) d\omega. \end{aligned} \quad (9)$$

The optimal excitation variances are obtained by differentiating (9) with respect to σ_x^2 and σ_w^2 and setting the partial derivatives to zero. This results in the following equations:

$$\begin{aligned} \int \frac{|a_w|^2 (P_y |a_x|^2 |a_w|^2 - \sigma_x^2 |a_w|^2 - \sigma_w^2 |a_x|^2)}{(\sigma_x^2 |a_w|^2 + \sigma_w^2 |a_x|^2)^2} d\omega &= 0, \\ \int \frac{|a_x|^2 (P_y |a_x|^2 |a_w|^2 - \sigma_x^2 |a_w|^2 - \sigma_w^2 |a_x|^2)}{(\sigma_x^2 |a_w|^2 + \sigma_w^2 |a_x|^2)^2} d\omega &= 0. \end{aligned} \quad (10)$$

One possible solution to (10) is when the codebooks contain entries such that $P_y(\omega) = \hat{P}_y(\omega)$ [3]. Thus the excitation variances are obtained by minimizing the spectral distance $d_{SD}(P_y, \hat{P}_y)$ between the modelled and observed noisy spectra resulting in

$$\begin{aligned} & \left[\begin{array}{cc} \left\| \frac{1}{|a_x(\omega)|^4} \right\| & \left\| \frac{1}{|a_x(\omega)|^2 |a_w(\omega)|^2} \right\| \\ \left\| \frac{1}{|a_x(\omega)|^2 |a_w(\omega)|^2} \right\| & \left\| \frac{1}{|a_w(\omega)|^4} \right\| \end{array} \right] \begin{bmatrix} \sigma_x^2 \\ \sigma_w^2 \end{bmatrix} \\ &= \left[\begin{array}{c} \left\| \frac{P_y(\omega)}{|a_x(\omega)|^2} \right\| \\ \left\| \frac{\hat{P}_y(\omega)}{|a_w(\omega)|^2} \right\| \end{array} \right]. \end{aligned} \quad (11)$$

The above equation set is different from the one obtained in [3] where the variances are obtained by minimizing $\|(P_y(\omega)|a_x(\omega)|^2|a_w(\omega)|^2 - \sigma_x^2|a_w(\omega)|^2 - \sigma_w^2|a_x(\omega)|^2)^2\|$, instead of $d_{SD}(P_y, \hat{P}_y)$.

The described log-likelihood approach has a mismatch between the criterion used in the variance estimation (spectral distance) and the criterion used in the global search (log-likelihood). The log-spectral distortion and Itakura-Saito distortion based estimators use the same criterion in both steps, thus avoiding this mismatch.

2.4. Envelope vs. periodogram

The variance estimation in the three methods described above is performed using symmetric distortion measures. Since the model provides the spectral envelope, to get a good fit between the observed and modelled spectra satisfying the assumption of small errors, it is necessary to use the envelope of the observed noisy speech instead of the periodogram i.e.,

$$P_y(\omega) = \frac{\sigma_y^2}{|a_y(\omega)|^2}, \quad a_y(\omega) = \sum_{k=0}^r a_{y_k} e^{-j\omega k}, \quad (12)$$

where a_{y_k} are the order- r AR-coefficients of the noisy signal and σ_y^2 is the corresponding excitation variance.

3. Iterative parameter estimation

An important feature shared by the proposed method and the method presented in [3] is the potential to perform well in the presence of non-stationary noise. The use of a noise codebook eliminates the need for noise estimation techniques, most of which perform well only for stationary noise. However, since estimation algorithms do work well for stationary noise, using estimated noise information could provide better performance than just using a-priori information. This is more so if the constraint on the noise codebook size results in a poor representation. In the following, we present an iterative scheme that attempts to combine the best features of both techniques. The iterative scheme also leads to significantly reduced complexity.

In the first step, an estimate of the spectral shape of noise is obtained using any noise estimation technique, for example the minimum statistics approach [2]. This estimate is used to search through the speech codebook to obtain the speech spectral shape that minimizes the relevant distortion measure. For example, for log-spectral distortion, the search involves calculating the excitation variances according to (6) and evaluating (4) for each speech codebook entry. The optimal speech spectral shape that results from this search is now used to find the best noise spectral shape from the noise codebook. The iterative procedure of alternatingly finding the optimal speech and

```

For each frame of noisy speech
i = 0
aw(i) = Estimated noise spectral shape
repeat
  i := i + 1
  Search speech CB with aw(i-1) to obtain ax(i), σx2(i), σw2(i)
  Search noise CB with ax(i) to obtain aw(i), σx2(i), σw2(i)
until convergence

```

Table 1: Iterative search algorithm

noise codebook entries and the associated variances continues until convergence, where convergence is said to occur when there is no improvement in the distortion measure used in the search. It may happen at the first step of the iteration that the estimated shape is better than all entries in the noise codebook. In this case, the iteration is terminated. Each iteration reduces the value of the chosen distortion measure. This, together with the fact that the codebooks are of finite size, guarantees convergence. In practice it might be useful to define an upper bound on the number of iterations to avoid a full search of the joint codebook. The complete algorithm is presented in table 1.

A large reduction in complexity results from not having a joint search through both codebooks. Consequently, it is possible to increase the size of the codebooks to provide better signal representation. It is possible for the iterative scheme to converge to a local optimum. This can be overcome to some extent by selecting at each stage the N best entries from the speech (noise) codebooks instead of just one. We refer to these N best entries as a subset.

4. Experimental results

To evaluate the performance of the proposed iterative method, experiments were conducted using ten utterances, five male and five female, from the TIMIT database. Neither the speakers nor the utterances were used in the training of the speech codebook. A 10-bit speech codebook (AR model order 10) and a 6-bit noise codebook (AR model order 6) were used in the experiments. The speech codebook was trained using the generalized Lloyd algorithm with 10 minutes of speech from the TIMIT database using the Itakura-Saito distortion measure [7]. The noise codebook was trained using white Gaussian noise, aircraft noise and vehicle noise from the Noisex-92 database.

Experiments were conducted for noisy speech at 5 dB and 10 dB input SNR for aircraft noise, (obtained from the Noisex-92 database), stationary white Gaussian noise, an artificially generated non-stationary white noise source (NS-1) and a real world non-stationary noise (NS-2). Noise data used in the experiments were not included in the training. The non-stationary white noise was generated by alternating the variance of white Gaussian noise every 500 ms between σ^2 and $10000\sigma^2$. A sample signal segment is shown in figure 2. The other non-stationary noise type was obtained by recording noise on a freeway as perceived by a pedestrian standing at a fixed point. A frame length of 240 samples with 50% overlap, with a Hann window was used in the codebook training and experiments. The LPC orders were 10, 6 and 16 for clean speech, noise and noisy speech respectively and the coefficients were obtained using the autocorrelation method. The subset sizes for the iterative method were experimentally determined and were fixed at

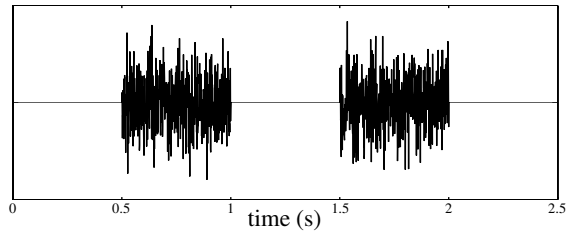


Figure 2: Artificially generated non-stationary white noise for $\sigma^2 = 1$.

10 entries for speech and 5 entries for noise. The Wiener filter was constructed as

$$H(\omega) = \frac{\sigma_x^2/|a_x(\omega)|^2}{\sigma_x^2/|a_x(\omega)|^2 + \sigma_w^2/|a_w(\omega)|^2}. \quad (13)$$

While in theory it is possible to iterate several times till the search complexity becomes equivalent to a full search of the joint codebook, in practice convergence was found to occur quite early in the iteration. For a subset size of one, the method converged after three iterations. For higher subset sizes such as three, convergence occurred after two iterations. This is intuitive since we expect convergence to occur earlier as the subset size is increased, since in the limit the subset size would be equal to the codebook size. To see the reduction in complexity due to the iterative method, let m, n denote the number of entries in the speech and noise codebooks respectively, let m_s, n_s be the cardinality of the speech and noise subsets, and let N denote the number of iterations. The non-iterative technique requires searching through mn combinations of codebook entries. The number of searches for the iterative method is $N(m_s n + m n_s) \approx N m n_s$ since the latter term dominates the sum. Thus there is a reduction in complexity provided $N n_s < n$. In the experiments, these values were $n_s = 5, n = 64$ and the average value of N was observed to be 2.

For performance comparisons, the following methods were implemented: the proposed log-spectral distortion (LS), Itakura-Saito distortion (IS) and log-likelihood (LL) based estimators, simple Wiener filtering and codebook constrained iterative Wiener filtering (CCIWF) [4]. For all methods, noise estimation was performed using the minimum statistics approach of [2]. The simple Wiener filter was constructed according to (13). The clean spectrum was obtained by subtracting the estimated noise spectrum from the noisy periodogram, followed by an LP-analysis to obtain the excitation variance and the AR-coefficients. The CCIWF method was implemented according to [4].

Tables 2 and 3 compare the performance of the different methods. The results shown in the tables were obtained by averaging the SNR results for each of the ten utterances. For stationary noise, the performance of the proposed estimators based on log-spectral distortion and Itakura-Saito distortion is comparable to the performance of CCIWF. This is to be expected since noise estimation works well for stationary noise. A possible reason for the relatively poor performance of the log-likelihood based estimator could be the mismatch in the criterion used for the variance estimation and the one used for the global search. The three proposed methods and CCIWF perform better than just using a Wiener filter in the enhancement. For both the non-stationary noise types, the proposed methods perform signifi-

cantly better than Wiener filtering and CCIWF. The latter two result in little or no gain in SNR for the non-stationary white noise and moderate gain for the freeway noise.

Noise Type	Wiener	CCIWF	LL	LS	IS
White	10.0	10.8	10.5	10.5	11.1
Aircraft	9.6	10.7	10.2	10.9	11.0
NS-1	5.1	5.1	8.8	9.6	10.3
NS-2	7.1	7.7	9.8	10.1	10.4

Table 2: Avg. SNR values for noisy speech at 5 dB input SNR

Noise Type	Wiener	CCIWF	LL	LS	IS
White	13.9	14.4	13.9	14.2	14.6
Aircraft	13.7	14.2	13.3	14.1	14.2
NS-1	10.2	10.1	12.1	12.9	13.5
NS-2	11.9	12.2	13.1	13.3	13.6

Table 3: Avg. SNR values for noisy speech at 10 dB input SNR

5. Conclusions

Using a-priori information about speech and noise results in improved speech enhancement, especially under non-stationary noise conditions where conventional methods fail. The results are consistent with the notion that using estimated noise information in addition to a-priori information leads to good performance. The iterative search technique presented in this paper avoids the high complexity associated with full search of the joint speech and noise codebook. Experiments show promising results and indicate that a-priori information based methods could be a topic for further research.

6. References

- [1] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 3, 2000.
- [2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.
- [3] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, 2001.
- [4] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 383–389, Sep 1996.
- [5] Y. Ephraim, "A minimum mean square error approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, pp. 829–832, 1990.
- [6] R. M. Gray, A. Buzo, A. H. Gray Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, pp. 367–376, Aug 1980.
- [7] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, vol. COM-28, pp. 84–95, Jan 1980.