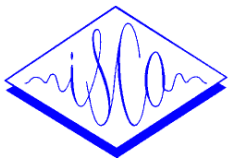

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ARDOSS: AUTOREGRESSIVE DOMAIN SPECTRAL SUBTRACTION FOR ROBUST SPEECH RECOGNITION IN ADDITIVE NOISE.

Hugo Van hamme

Lernout & Hauspie Speech Products N.V.
Koning Albert I laan 64
B-1780 Wemmel, BELGIUM

ABSTRACT

The first and second order statistics of the LPC parameters of speech corrupted by additive noise are predicted based on the first few lags of the autocorrelation of the noise. The computed mean allows a correction on the LPC parameters without reference to an assumed state and for any type of HMM emission models. This mean is equivalent to a 5 dB noise suppression. Additional robustness is gained when the predicted covariance in the AR-domain is transposed to the cepstral domain to correct the emission probabilities in a single-Gaussian HMM. These conclusions are drawn from speaker-dependent experiments on the NOISEX-92 database.

For a p-th order LPC analysis, correction of the mean is accomplished in $O(p^2)$ floating point operations (flops). The full covariance correction requires $O(p^3)$ flops. An $O(p^2)$-approximation that yields comparable performance in practice is given.

I. INTRODUCTION

Corruption of clean speech by additive noise changes the observed spectrum and hence deteriorates the matching of the resulting acoustic vectors with their speech models trained on clean data. Many of the few available low-cost methods for compensation of LPC-cepstra against this phenomenon are covered in [1]. In [2], the noise autocorrelation is subtracted from the lags of the noisy signal, but this frequently yields non-positive definite or even singular normal equations and unstable prediction polynomials, which makes the subsequent recursive computation of cepstra [3] incorrect.

This paper presents two low-cost corrections: the first is an approximation to linear spectral subtraction, working directly on LPC vectors. It can be applied to any type of speech modelling based on LPC-derived acoustic vectors, including the fast discrete-density HMM. The second correction takes the reduced reliability of noise-corrupted acoustic vectors into account by modifying the emission densities of HMM models, more specifically the covariance of single-Gaussians. Other authors [4] have obtained good results by applying covariance corrections.

II. FIRST ORDER PERTURBATION ANALYSIS OF LPC PARAMETERS.

Let s_t, n_t and x_t denote samples of the noise-free speech, the stationary noise and the noisy speech respectively: $x_t = s_t + n_t$ where $t = -\infty \ldots \infty$ and only x_t is observable. Its autocorrelation of order i, $r_{x,i}$, is estimated from a frame of N samples as:

$$r_{x,i} = \frac{1}{N}\sum_{t=i}^{N-1}(w_t x_t)(w_{t-i}x_{t-i}) \qquad i = 0 \ldots p \tag{1}$$

where w_t is a window. Similarly, $r_{s,i}$ and $r_{n,i}$ denote the windowed autocorrelation estimate of s_t and n_t respectively.

For noise-free data, the p-th order LPC parameters $\boldsymbol{a}_s = [a_{s,1} \;\; a_{s,2} \;\; \ldots \;\; a_{s,p}]^t$ (superscript t denotes matrix transpose) are obtained in $O(p^2)$ flops using the Durbin [7, 3] algorithm as:

$$\boldsymbol{a}_s = -\boldsymbol{R}_s^{-1}[r_{s,1} \;\; r_{s,2} \;\; \ldots \;\; r_{s,p}]^t \tag{2}$$

where $\boldsymbol{R}_s$ is a symmetrical Toeplitz matrix:

$$\boldsymbol{R}_s = \begin{bmatrix} r_{s,0} & r_{s,1} & \cdots & r_{s,p-1} \\ r_{s,1} & r_{s,0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{s,1} \\ r_{s,p-1} & \cdots & r_{s,1} & r_{s,0} \end{bmatrix} \tag{3}$$

In the presence of noise, $r_{s,i}$ is not available, but if the speech and the noise are independent, it can be estimated from $r_{x,i}$ by subtraction of a noise autocorrelation estimate. This is the time-domain analogue of the linear spectral subtraction method. Unfortunately, the positive definiteness of the Toeplitz matrix (3) and the stability of the prediction polynomial are easily lost and additional measures must be taken [2]. In order to avoid such interventions, the noise correction will be done on the prediction polynomial, based on a linearised perturbation analysis.

Assume that $r_{s,i}$ in (2) is perturbed by a stochastic component $\Delta r_{s,i}$ which is caused by the additive noise. The speech signal is considered as deterministic in the following analysis. The perturbation of $\boldsymbol{a}_s$ is then approximately:

$$\Delta \boldsymbol{a}_s = \sum_{i=0}^{p} \frac{\partial \boldsymbol{a}_s}{\partial r_{s,i}} \Delta r_{s,i} \tag{4}$$

Algebraic manipulation yields

$$\Delta \boldsymbol{a}_s = -\boldsymbol{R}_s^{-1}\boldsymbol{B}_s \, \Delta \boldsymbol{r}_s$$

where $\boldsymbol{B}_s$ is a p-by-p+1 Hankel-plus-(nearly)-Toeplitz matrix:

$$\boldsymbol{B}_s = \begin{bmatrix} a_{s,1} & a_{s,2} & \cdots & a_{s,p-1} & a_{s,p} & 0 \\ a_{s,2} & \iddots & \iddots & a_{s,p} & 0 & 0 \\ \vdots & a_{s,p-1} & \iddots & \iddots & \iddots & 0 \\ a_{s,p-1} & a_{s,p} & 0 & \iddots & \iddots & 0 \\ a_{s,p} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & a_{s,1} & 1 & \ddots & 0 & 0 \\ 0 & a_{s,2} & a_{s,1} & \ddots & 0 & 0 \\ \vdots & & \ddots & \ddots & 1 & 0 \\ 0 & a_{s,p-1} & a_{s,p-2} & \cdots & a_{s,1} & 1 \end{bmatrix}$$

and $\Delta \boldsymbol{r}_s = [\Delta r_{s,0} \quad \Delta r_{s,1} \quad \cdots \quad \Delta r_{s,p}]'$.

The first- and second order statistics of the perturbation of the LPC parameters are now easily found:

$$E\{\Delta \boldsymbol{a}_s\} = -\boldsymbol{R}_s^{-1} \boldsymbol{B}_s E\{\Delta \boldsymbol{r}_s\} \tag{5}$$

and $Cov\{\Delta \boldsymbol{a}_s\} = \boldsymbol{R}_s^{-1} \boldsymbol{B}_s Cov\{\Delta \boldsymbol{r}_s\} \boldsymbol{B}_s' \boldsymbol{R}_s^{-1}$

For high SNR, $\boldsymbol{R}_s$ and $\boldsymbol{B}_s$ can successfully be approximated by $\boldsymbol{R}_x$ and $\boldsymbol{B}_x$ in the above. Hence, for every frame, the desired $\boldsymbol{a}_s$ can be estimated as:

$$\hat{\boldsymbol{a}}_s = \boldsymbol{a}_x + \boldsymbol{R}_x^{-1} \boldsymbol{B}_x E\{\Delta \boldsymbol{r}_s\} \tag{6}$$

with covariance

$$Cov\{\hat{\boldsymbol{a}}_s\} \approx \boldsymbol{R}_x^{-1} \boldsymbol{B}_x Cov\{\boldsymbol{r}_x\} \boldsymbol{B}_x' \boldsymbol{R}_x^{-1} \tag{7}$$

This covariance will later be transposed to the cepstral covariance, which is used to adapt the HMM. The approximation of $\boldsymbol{R}_s$ and $\boldsymbol{B}_s$ by $\boldsymbol{R}_x$ and $\boldsymbol{B}_x$ will be most troublesome in silence. However, in that case, the linear approximation (4) will be bad any way. For example, it causes the effect that when using $\boldsymbol{R}_s$, the norm of the LPC correction becomes infinite in silence, which is wrong. Substitution of $\boldsymbol{R}_s$ by $\boldsymbol{R}_x$ limits the damage.

As in cepstral normalisation [5], the LPC compensation (6) is additive and SNR-dependent, although no explicit measurement of the SNR is done. In the present method, the noise correction operates in the AR-domain, rather than in the cepstral domain. Another valid comparison is with spectral subtraction, where the noise spectrum is subtracted from the noisy speech spectrum. Here, the effect of subtracting the noise autocorrelation is done in the AR-domain after linearisation, hence the name AutoRegressive DOmain Spectral Subtraction (ARDOSS). For those SNR where the first order Taylor series is valid, the mean compensation (6) is expected to yield a robustness similar to that offered by spectral subtraction, without resorting to FFT computations. A correction of the covariance based on a noise model, as in (7), was also applied by Gales [4], but for different acoustic vectors.

III. STATISTICS OF THE AUTOCORRELATION.

In order to finalise (6) and (7), the mean and covariance of (1), conditioned on the speech, will now be computed assuming that $\{n_t\}$:

1. is independent of the speech,
2. is wide-sense stationary with autocorrelation c_i,
3. has zero mean,
4. has zero 3rd and 4th order cumulants (e.g. Gaussian).

Therefore, rewrite (1) as:

$$r_{x,i} = r_{s,i} + \tfrac{1}{N} \sum_{t=i}^{N-1} (w_t n_t)(w_{t-i} s_{t-i}) + \tfrac{1}{N} \sum_{t=i}^{N-1} (w_t s_t)(w_{t-i} n_{t-i}) + \tfrac{1}{N} \sum_{t=i}^{N-1} (w_t n_t)(w_{t-i} n_{t-i})$$

Due to the assumptions 1 and 3, the mean of the second and the third term vanish. Assumption 2 then yields:

$$E\{r_{x,i}\} = r_{s,i} + E\{r_{n,i}\} \tag{8}$$

where $E\{r_{n,i}\}$ can be estimated by averaging multiple observations of $r_{n,i}$ or by estimating c_i from a long noise record and using

$$E\{r_{n,i}\} = c_i \tfrac{1}{N} \sum_{t=i}^{N-1} w_t w_{t-i}$$

Eq. (8) provides the mean of the perturbed autocorrelations, so $E\{\Delta \boldsymbol{r}_s\} = E\{\boldsymbol{r}_n\}$ in (6).

The covariance of the autocorrelation is found applying the bilinearity of the covariance operator, the theorem of Leonov and Shiryaev [6] and recalling that the cumulants of order 3 and 4 are assumed to be zero:

$$\begin{aligned} &Cov\{r_{x,m}, r_{x,n}\} \\ &= \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} Cov\{(w_t x_t)(w_{t-m} x_{t-m}), (w_k x_k)(w_{k-n} x_{k-n})\} \\ &= \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k} \tilde{s}_{t-m} \tilde{s}_{k-n} + \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t-m,k-n} \tilde{s}_t \tilde{s}_k \\ &+ \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k-n} \tilde{s}_{t-m} \tilde{s}_k + \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t-m,k} \tilde{s}_t \tilde{s}_{k-n} \\ &+ \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k} \tilde{c}_{t-m,k-n} + \tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k-n} \tilde{c}_{t-m,k} \end{aligned}$$

where $\tilde{c}_{k,l} = w_k w_l c_{k-l}$ and $\tilde{s}_k = w_k s_k$.

Although a proof of the following property is beyond the scope of this paper, one can show that asymptotically:

$$\begin{aligned} &\tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k} \tilde{s}_{t-m} \tilde{s}_{k-n} \xrightarrow{N \to \infty} \frac{g_{m-n}}{N} + O(N^{-2}) \\ &\tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t-m,k-n} \tilde{s}_t \tilde{s}_k \xrightarrow{N \to \infty} \frac{g_{m-n}}{N} + O(N^{-2}) \\ &\tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k-n} \tilde{s}_{t-m} \tilde{s}_k \xrightarrow{N \to \infty} \frac{g_{m+n}}{N} + O(N^{-2}) \\ &\tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t-m,k} \tilde{s}_t \tilde{s}_{k-n} \xrightarrow{N \to \infty} \frac{g_{m+n}}{N} + O(N^{-2}) \\ &\tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k} \tilde{c}_{t-m,k-n} \xrightarrow{N \to \infty} \frac{h_{m-n}}{N} + O(N^{-2}) \\ &\tfrac{1}{N^2} \sum_{t=m}^{N-1} \sum_{k=n}^{N-1} \tilde{c}_{t,k-n} \tilde{c}_{t-m,k} \xrightarrow{N \to \infty} \frac{h_{m-n}}{N} + O(N^{-2}) \end{aligned}$$

with $g_k = \tfrac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_s(\omega) \varphi_n(\omega) e^{j2\pi k} d\omega$

and $h_k = \tfrac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_n^2(\omega) e^{j2\pi k} d\omega$

where φ_s and φ_n denote the power spectrum of the speech (assuming stationarity !) and the noise respectively. Combining the equations above, one obtains:

$$Cov\{\boldsymbol{r}_x\} \approx \tfrac{2}{N}(\boldsymbol{F}_{toep} + \boldsymbol{F}_{hank} - \boldsymbol{H}_{hank}) \qquad (9)$$

where

$$\boldsymbol{F}_{toep} = \begin{bmatrix} f_0 & f_1 & & f_p \\ f_1 & f_0 & \ddots & \\ & \ddots & \ddots & f_1 \\ f_p & & f_1 & f_0 \end{bmatrix}, \quad \boldsymbol{F}_{hank} = \begin{bmatrix} f_0 & f_1 & & f_p \\ f_1 & \iddots & \iddots & f_{p+1} \\ & \iddots & \iddots & \\ f_p & f_{p+1} & & f_{2p} \end{bmatrix},$$

$\boldsymbol{H}_{hank}$ is defined from h_0 through h_{2p} in the same way $\boldsymbol{F}_{hank}$ is defined from f_0 through f_{2p}, and

$$f_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \varphi_x(\omega)\varphi_n(\omega)e^{j2\pi k}d\omega \qquad (10)$$

Equation (10) can be evaluated using the spectra φ_x and φ_n, which can be readily estimated from the available data. Equations (7) and (9) provide all information for computing the covariance due to noise on the corrected LPC parameters.

Notice that all the information regarding the noise model is derived from an autocorrelation measurement only. There is no need to measure the covariance of the acoustic vectors of the noise as in [4].

Evaluation of (9) requires the computation of the autocorrelation $r_{ab,i}$ of a process with known power spectrum $\varphi_a\varphi_b$. Since these are noise or noisy speech spectra, it is reasonable to assume that an AR-model is available:

$$\varphi_a(\omega) = \frac{\sigma_a^2}{A(e^{j\omega})A(e^{-j\omega})} \text{ and } \varphi_b(\omega) = \frac{\sigma_b^2}{B(e^{j\omega})B(e^{-j\omega})}$$

where $A(.)$ and $B(.)$ are polynomials with unit constant term of degree $n_a \le p$ and $n_b \le p$ respectively. To find the autocorrelation, first $A(.)$ and $B(.)$ are multiplied in $O(n_a n_b)$ flops to yield a prediction polynomial of degree n_a+n_b. Subsequently, the prediction polynomials down to the first order are constructed by *reverse* Durbin recursion [3] in $O((n_a+n_b)^2)$ flops. Then, applying the Yule-Walker equations [7, 3], the autocorrelation lags can be obtained recursively in $O((n_a+n_b)^2)$ flops.

IV. COVARIANCE OF THE LPC-CEPSTRA.

The RPS-scaled negative LPC-cepstra b_k are obtained from the LPC parameters a_k using the recursion:

$$b_1 = a_1 \qquad (11)$$

and

$$b_n = na_n - \sum_{k=1}^{n-1} b_k a_{n-k} \qquad n = 2 \ldots p \qquad (12)$$

This recursion is applied to the corrected LPC parameters (6). Neglecting third order cumulants and applying [6], recursion (12) can be converted into a recursion on the covariances. One thus obtains:

$$\begin{aligned} Cov\{b_1, a_m\} &= Cov\{a_1, a_m\} \qquad m = 1 \ldots p \\ Cov\{b_1, b_1\} &= Cov\{a_1, a_1\} \\ Cov\{b_n, a_m\} &= n\,Cov\{a_n, a_m\} - \sum_{k=1}^{n-1} Cov\{b_k, a_m\}a_{n-k} \\ &\quad - \sum_{k=1}^{n-1} Cov\{a_k, a_m\}b_{n-k} \qquad m = 1 \ldots p \end{aligned} \qquad (13)$$

$$\begin{aligned} Cov\{b_n, b_m\} &= n\,Cov\{b_m, a_n\} - \sum_{k=1}^{n-1} Cov\{b_k, b_m\}a_{n-k} \\ &\quad - \sum_{k=1}^{n-1} Cov\{b_m, a_k\}b_{n-k} \qquad m = 1 \ldots p \end{aligned} \qquad (14)$$

The above recursion requires $O(p^3)$ computations. Therefore, §V provides an $O(p^2)$ approximation to the cepstral covariance for a particular, but important, case.

V. FAST CEPSTRAL CORRECTION

The computational load of the full covariance correction is rather high, but not unfeasible. In some applications, one will prefer to approximate the cepstral noise covariance such that it can be computed in $O(p^2)$ flops. Moreover, the recognition experiments below show a comparable performance.

When working with HMMs with (multi-)Gaussian emission probabilities, it is sufficient to know the diagonal entries of the cepstral covariance: $Var\{b_n\} = Cov\{b_n, b_n\}$. In view of (11), there is no need to execute recursions (13) and (14), if one assumes that for a given set-up, $Var\{b_n\}/Var\{b_1\}$ depends on n only. Then the cepstral covariance can be found from $Var\{a_1\}$ in $O(p)$ flops. Currently, the same ratios as the average state covariance of the HMM are used. With this assumption, it is sufficient to compute the upper left entry of $Cov\{\hat{\boldsymbol{a}}_s\}$ in (7). The computation of the LPC correction and the first row of $\boldsymbol{R}_x^{-1}$ is done simultaneously in $O(p^2)$ flops using the Levinson algorithm [7].

VI. EXPERIMENTS

Tests were performed on the NOISEX-92 database [8] with noise statistics derived from a separate noise recording of 5000 samples. The recogniser is a diagonal-covariance single-Gaussian speaker-dependent HMM system used to recognise *isolated* digits. The feature vectors are cepstra derived from the (corrected) LPC's, their first order difference, the logarithm of the energy and its first order difference. Each word is characterised by 7 HMM states which share a common diagonal covariance matrix per word, preceded and followed by a single-state silence model with its own covariance matrix.

The noisy speech signal is filtered by a 300Hz-3000Hz band pass filter (high pass filtering for CAR noise), pre-emphasised and cut into 30 ms windows with 20 ms overlap. A Hamming window is applied, followed by 10th order LPC analysis and a 12th order cepstrum recursion.

The recognition performance using four correction methods is shown in Figure 1 through Figure 3. The correction methods are:

- 'none': uses uncorrected LPC-cepstra,
- 'mean': uses corrected LPC-cepstra,
- 'fast': as 'mean' plus the $O(p^2)$ covariance correction,
- 'full': as 'mean' plus the $O(p^3)$ covariance correction.

The noise AR-models are of order 10 to catch the steep slopes in the noise spectrum due to the pre-filtering.

Without pre-filter, an order of 5 is sufficient.

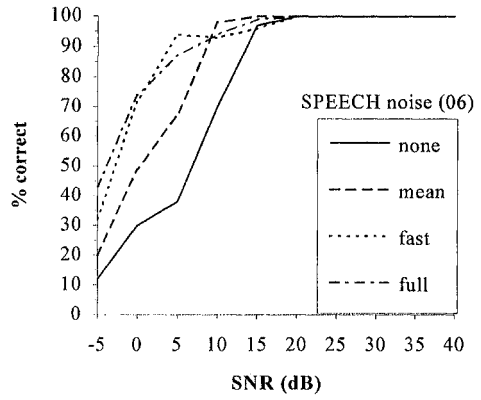

Figure 1: noise-sensitivity for telephone-bandwidth speech in SPEECH noise.

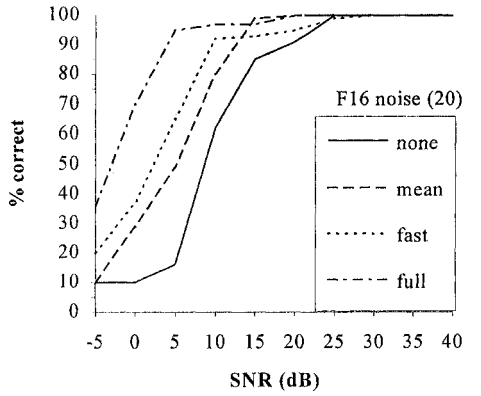

Figure 2: noise-sensitivity for telephone-bandwidth speech in F16 noise.

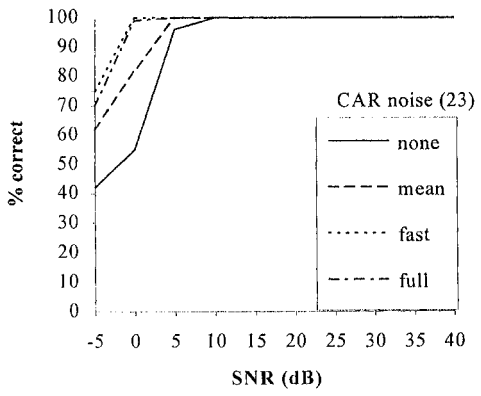

Figure 3: noise-sensitivity for high pass (300 Hz) filtered speech in CAR noise.

In all cases, the AR-domain correction improves the robustness against additive noise. It must be stressed that this may not be the case for non-stationary noise.

The covariance correction offers further robustness. In time-critical implementations, the 'fast' approximation suffices.

VII. CONCLUSION

The presented LPC compensation has the computational cost of a Levinson recursion and performs statistically like spectral subtraction. Improved performance is obtained when covariance correction is implemented. The $O(p^2)$ covariance correction approaches its $O(p^3)$ version in performance.

VIII.REFERENCES

[1] Acero A. (1993). *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers.

[2] C.K. Un and K.Y. Choi (1981). Improving LPC Analysis of Noisy Speech by Autocorrelation Subtraction Method. *ICASSP 81*, pp. 1082-1085.

[3] L.R. Rabiner and R.W. Schafer (1978). *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[4] Gales M.J.F. and S.J. Young (1993). Cepstral Parameter Compensation for HMM Recognition in Noise. *Speech Communication 12 (1993)*, pp. 231-239.

[5] Acero A. and R.M. Stern (1990). Environmental Robustness in Automatic Speech Recognition. *ICASSP 90*, pp. 849-852.

[6] Leonov V.P. and A.N. Shiryaev (1959). On a Method of Calculation of Semi-Invariants, *Theory of Probability and Applications,* Vol. IV, No. 3, 1959, pp. 319-329.

[7] G.H. Golub and C.F. Van Loan (1989). *Matrix Computations, second edition.* The Johns Hopkins University Press, Baltimore, 1989.

[8] Varga A. , H.J.M. Steeneken, M. Tomlinson and D. Jones (1992). *The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition.*