Enhanced Speech Features by Single-Channel Joint Compensation of Noise and Reverberation

Author : Matthias Wölfel

Professor:陳嘉平

Reporter:葉佳璋

Outline

- Introduction
- Speech Feature Enhancement By Particle Filters
- Evaluation Of Samples
- Prediction Of Samples
- Distortion Compensation
- Late Reverberation Estimation
- Putting The Pieces Together
- Experiment

Introduction

- A lot of research and development are devoted to address one of the two distortions, namely additive noise or reverberation.

- We observe that a simple concatenation of techniques addressing either additive noise or reverberation.

Introduction

- Method assume that the reverberant power spectrum r_k is a scaled or weighted summation over previous frame

$$x_k^{(reverberant)} = x_k + r_k = x_k + \sum_{m=1}^{M} s_m x_{k-m}$$

- The scale terms can be determined by the Rayleigh distribution and adjust by an estimate of the reverberation.

Speech Feature Enhancement By Particle Filters

- Speech feature enhancement techniques on nonstationary distortions, can be formulated as a tracking problem.
- The clean speech features x_k have to be estimated for each frame k, given the current observation and history of the noisy feature $y_{1:k}$.

Speech Feature Enhancement By Particle Filters

- A general description of such a system that relates two stochastic process
  - State $(X_k)_{k \in N}$: representing a hidden, inner system.
  - $(Y_k)_{k \in N}$:corresponding observation of measurement.
- In there most general (discrete) form are as
  - The state equation $x_k = f(x_{k-1}, v_{k-1})$
  - The observation equation $y_k = g(x_k, w_k)$

◦ f and g :the nonlinear transition and observation function

◦ x_k and y_k : the state and observation vector

◦ v_k and w_k : the process noise and measurement noise

Speech Feature Enhancement By Particle Filters

- The state equation characterizes the state transition probability $p(x_k \mid x_{k-1})$, while the observation equation describe the probability $p(y_k \mid x_{k-1})$ which is coupled to the measurement noise model.
- The minimum mean square error(MMSE) solution to a tracking problem, which relates x and y by the probabilistic relationship $p(x_k \mid y_{1:k})$

$$E\{x_k \mid y_{1:k}\} = \int x_k \, p(x_k \mid y_{1:k}) dx$$

Tracking the Individual Distortion Types

- We aim to decompose the observed signal y into three parts:
  - The energy of the clean signal x
  - The energy cause by additive noise a
  - The energy caused by reverberation r
- We do not tracking the impulse response or late reverberation, but the difference to an energy estimate of reverberation.

Tracking the Individual Distortion Types

- Tracking of the additive noise a_k and scale term s_k, instead of signal x_k given the distorted observation y_k

$$p(x_k \mid y_{1:k}) = \int\int p(x_k, a_k, s_k \mid y_{1:k})\, da_k ds_k$$

$$p(x_k, a_k, s_k \mid y_{1:k}) = p(x_k \mid y_{1:k}, a_k, s_k)\, p(a_k, s_k \mid y_{1:k})$$

Change in integration order, we obtain

$$E\{x_k \mid y_{1:k}\} = \int\int v(y_{1:k}, a_k, s_k)\, p(a_k, s_k \mid y_{1:k})\, da_k ds_k$$

$$v(y_{1:k}, a_k, s_k) = \int x_k\, p(x_k \mid y_{1:k}, a_k, s_k)\, dx_k$$

Tracking the Individual Distortion Types

- Folding two vectors into one super vector $d = \begin{bmatrix} a \\ s \end{bmatrix}$

$$p(d_k \mid y_{1:k}) = p(d_k \mid y_k, y_{1:k-1}) = \frac{p(y_k \mid d_k, y_{1:k-1})\, p(d_k \mid y_{1:k-1})}{p(y_k \mid y_{1:k-1})} = \frac{p(y_k \mid d_k)\, p(d_k \mid y_{1:k-1})}{p(y_k \mid y_{1:k-1})}$$

Which can be rewrite by Chapman-Kolmogorov equation as

$$p(d_k \mid y_{1:k-1}) = \int p(d_k \mid d_{k-1})\, p(d_{k-1} \mid y_{1:k-1})\, dd_{k-1}$$

The normalize term can solved by

$$p(y_k \mid y_{1:k-1}) = \int p(d_k, y_k \mid y_{1:k-1})\, dd_k = \int p(d_k \mid y_{1:k-1})\, p(y_k \mid d_k)\, dd_k$$

- Method to model the transition probability

$$p(d_k \mid d_{k-1}) = \begin{bmatrix} p(a_k \mid a_{k-1}) \\ p(s_k \mid s_{k-1}) \end{bmatrix}$$

Monte Carlo Sampling

- We aim to approximate the posterior density by weighted approximation as

$$p\left(d_k \mid y_{1:k}\right) \approx \sum_{s=1}^{S} \tilde{w}_k^{(s)} \delta(d_k - d_k^{(s)})$$

$$\frac{p\left(y_k \mid d_k\right) p\left(d_k \mid y_{1:k-1}\right)}{p\left(y_k \mid y_{1:k-1}\right)} = \frac{p\left(d_k, y_k \mid y_{1:k-1}\right)}{p\left(y_k \mid y_{1:k-1}\right)}$$

$$p\left(d_k, y_k \mid y_{1:k-1}\right) \approx \frac{1}{S} \sum_{s=1}^{S} P\left(d_k^{(s)} \mid d_{k-1}^{(s)}\right) P\left(y_k \mid d_k^{(s)}\right) \qquad p\left(y_k \mid y_{1:k-1}\right) \approx \frac{1}{S} \sum_{s=1}^{S} P\left(y_k \mid d_k^{(s)}\right)$$

Where the weight $w_k^{(s)} = P\left(y_k \mid d_k^{(s)}\right)$ are represented by the corresponding likelihood for each sample s out of S samples.

- Those samples are known as particles and the filter process is called particle filter.

Evaluation Of Samples

- The relation can be approximate by

$$x = y + \ln\left(1 - e^{n-y}\right) + e_{\theta} + e_{envelope} \approx \ln\left(e^{y} - e^{n}\right) \quad n = u\left(d, r\right) = u\left(a, s, r\right)$$

- The error term

$$e_{\theta}\left(\Omega\right) = \ln\left(1 + \frac{2\cos\theta\left(\Omega\right)}{\cosh\left\{\ln\left|N\left(\Omega\right)\right| - \left\{\ln\left|X\left(\Omega\right)\right|\right\}\right\}}\right)$$

- The average value is close to zero and that $\theta(\Omega)$ is Gaussian distributed.
- In the case of cepstral or spectral envelope techniques, a second error term $e_{envelope}$ is further weakened.

Weight Calculation For Each Sample

- To evaluate each sample $n_k = u(d_k, r_k)$ according to the likelihood function

$$p\left(y_k \mid d_k^{(s)}\right) = \frac{p_{speech}\left(y_k + \ln\left(1 - e^{n_k^{(s)} - y_k}\right)\right)}{\Pi_{b=1}^{B}\left|1 - e^{n_{k,b}^{(s)} - y_{k,b}}\right|}$$

- $p_{speech}(\cdot)$ denote the prior speech density represented by a Gaussian mixture model which has been trained by clean speech.

Weight Calculation For Each Sample

- To get the normalize weights, the likelihoods have to be divided by the sum over likelihoods.

$$\tilde{w}_k^{(s)} = \frac{p\left(y_k \mid d_k^{(s)}\right)}{\sum_{m=1}^{S} p\left(y_k \mid d_k^{(m)}\right)}$$

- Note that the normalize weight can only be evaluated if $n_{k,b}^{(s)} < y_{k,b} \forall b \in B$.
- If this constraint is not satisfied, it has to be rejected by setting the particle weight to zero.

Prediction Of Samples

- Tracking requires the prediction of the distortion d_k given the previous estimate d_{k-1}.
- The simplest way to model the evolution of distortions is a random walk

$$a_k = a_{k-1} + \varepsilon_k$$

- a_k could represent the noise spectrum estimate, while the random term $\varepsilon_k \sim N\left(0, \Sigma^{random}\right)$

Predicted Walk By Static Autoregressive Processes

- To use an autoregressive process $A^{(1:L)}$, where L denotes the order, to predict the evolution of additive noise

$$a_k = A^{(1)} a_{k-1} + A^{(2)} a_{k-2} + \ldots + A^{(L)} a_{k-m} + \varepsilon_k = A^{(1:L)} a_{k-1:k-L} + \varepsilon_k$$

- Two components that have to be learned
  - ➢ The linear prediction matrix $A^{(1:L)}$
  - ➢ The covariance matrix $\Sigma^{AR} = diag(E\{\varepsilon\varepsilon^T\})$

Predicted Walk By Static Autoregressive Processes

- Minimization of the prediction error norm

$$A^{(1:L)} = E\left\{ a_k a_{k-1:k-L}^T \right\} E\left\{ a_{k-1:k-L} a_{k-1:k-L}^T \right\}^{-1}$$

$$E\left\{ a_k a_{k-1:k-L}^T \right\} = \frac{1}{K} \sum_{k=1}^{K} a_k a_{k-1:k-L}^T$$

$$E\left\{ a_{k-1:k-L} a_{k-1:k-L}^T \right\} = \frac{1}{K} \sum_{k=1}^{K} a_{k-1:k-L} a_{k-1:k-L}^T$$

- The static sample covariance matrix can then be calculated by

$$\Sigma^{AR} = \frac{1}{K} \sum_{k=1}^{K} \left(a_k - A^{(1:L)} a_{k-1:k-L} \right) \times \left(a_k - A^{(1:L)} a_{k-1:k-L} \right)^T$$

Predicted Walk by Dynamic Autoregressive Process

- In order to cope with changing environments, this requires an integrated estimate

$$a_k = A_{k-1} a_{k-1} + \varepsilon_k$$

$$A_k = A_k^1 = E\left\{a_k a_{k-1}^T\right\} E\left\{a_{k-1} a_{k-1}^T\right\}^{-1}$$

- Sum over all particle s=1,2, …, S for the current $a_k^{(s)}$ and previous $a_{k-1}^{(s)}$ noise estimate.

$$E\left\{a_k a_{k-1}^T\right\} = \frac{1}{S} \sum_{s=1}^{S} p\left(y_k \mid a_k^{(S)}\right) a_k^{(S)} a_{k-1}^{(S)T}$$

$$E\left\{a_{k-1} a_{k-1}^T\right\} = \frac{1}{S} \sum_{s=1}^{S} p\left(y_k \mid a_k^{(s)}\right) a_{k-1}^{(s)} a_{k-1}^{(s)T}$$

$$\hat{\Sigma}_k^{AR} = \sum_{s=1}^{S} \tilde{w}_k^{(s)} \left(a_k^{(s)} - A_k a_{k-1}^{(s)}\right) \times \left(a_k^{(s)} - A_k a_{k-1}^{(s)}\right)^T$$

Distortion Compensation

- To solve for the nonlinear relation $y \approx \ln\left(1 + e^{n_k - x_k}\right)$ will present next.
- Vector Taylor series(VTS) expansion around the oth Gaussian's mean μ_o .

$$p\left(x_k \mid y_{1:k}, n_k\right) = \sum_{o=1}^{O} p\left(x_k, o \mid y_{1:k}, n_k\right)$$

$$p\left(x_k, o \mid y_{1:k}, n_k\right) = p\left(o \mid y_{1:k}, n_k\right) p\left(x_k \mid o, y_{1:k}, n_k\right)$$

- We can pull the sum over o out of the integral

$$v\left(y_{1:k}, a_k, s_k\right) = \int x_k \, p\left(x_k \mid y_{1:k}, a_k, s_k\right) dx_k$$

$$v\left(y_{1:k}, d_k\right)^{(VTS)} = \sum_{o=1}^{O} p\left(o \mid y_{1:k}, n_k\right) \int x_k \, p\left(x_k \mid o, y_{1:k}, n_k\right) dx_k$$

Gaussian Mixture Approach

- The effect of n_k to the oth Gaussian in the log spectral domain is

$$e^{\mu_o^{'}} = e^{\mu_o} + e^{n_k}$$

$$\mu_o^{'} = \mu_o + \underbrace{\ln\left(1 + e^{n_k - \mu_o}\right)}_{\Delta\mu_o, n_k}$$

- Instead of shifting the mean, we can shift the corrupted spectrum in the opposite direction to obtain

$$x_k = y_k - \Delta_{\mu_o, n_k}$$

Gaussian Mixture Approach

- This deterministic relationship yields

$$p\left(x_k \mid o, y_{1:k}, n_k\right) = \delta_{y_k - \Delta\mu_o, n_k}$$

$$v^{(GMA)}\left(y_{1:k}, n_k\right)$$

$$= \sum_{o=1}^{O} p\left(o \mid y_{1:k}, n_k\right) \int x_k \delta_{y_k - \Delta_{\mu_o, n_k}} dx_k$$

$$= \sum_{o=1}^{O} p\left(o \mid y_{1:k}, n_k\right)\left(y_k - \Delta_{\mu_o, n_k}\right)$$

$$= y_k - \sum_{o=1}^{O} p\left(o \mid y_{1:k}, n_k\right) \Delta_{\mu_o, n_k}$$

Statistical Inference Approach

- Use the relationship from

$$x = y + \ln\left(1 - e^{n-y}\right)$$

- The marginal density $p\left(x_k \mid y_{1:k}, n_k\right)$ becomes deterministic

$$p\left(x_k \mid y_{1:k}, n_k\right) = \delta_{y_k + \ln\left(1 - e^{n_k - y_k}\right)}\left(x_k\right)$$

$$\begin{aligned} & v^{(SIA)}\left(y_{1:k}, n_k\right) \\ & = \int x_k \delta_{y_k + \ln\left(1 - e^{n_k - y_k}\right)}\left(x_k\right) dx_k \\ & = y_k + \ln\left(1 - e^{n_k - y_k}\right) \end{aligned}$$

Multistep Linear Prediction Estimation of Late Reverberation

- In order to estimate the correlation, it has been proposed to use MSLP

$$y[n] = \sum_{m=1}^{M} c_m y[n-m-D] + e[n]$$

- The solution for the MSLP coefficients

$$c = \left(E\left\{ y[n-D]\, y[n-D]^T \right\} \right)^{-1} E\left\{ y[n-D]\, y[n]^T \right\}$$

- An estimate of the sequence r[n] can be obtained by the observe sequence with MSLP

$$r[n] = \sum_{m=1}^{M} c_m y[n-m-D]$$

Particle Initialization

- The prior distortion density

$$p(d_0) = \begin{bmatrix} p(a_0) \\ p(s_0) \end{bmatrix}$$

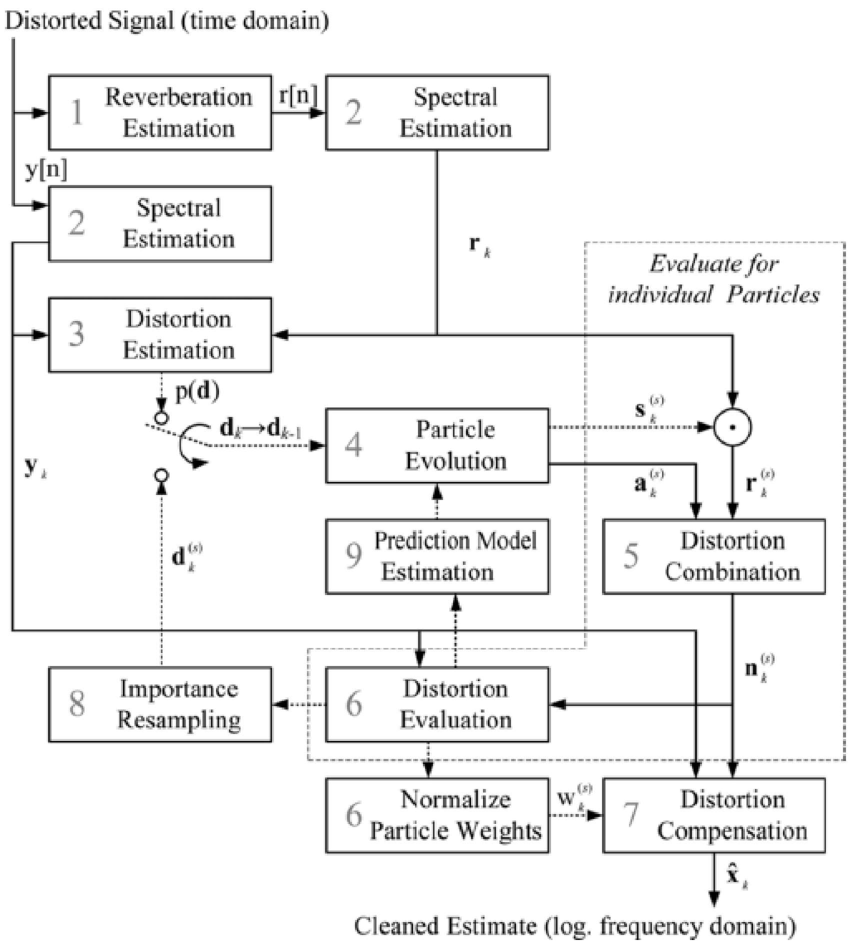
  - Prior overall distortion density $p(n_0) = N(\mu_n, \Sigma_n)$
  - Prior reverberation density $p(r_0) = N(\mu_r, \Sigma_r)$

$$p(a_0) = N(\mu_a, \Sigma_n) \qquad \mu_a = \ln\left(e^{\mu_n} - e^{\mu_r}\right)$$

- The prior scale density $p(s_0) = N(\mu_s, \Sigma_s)$

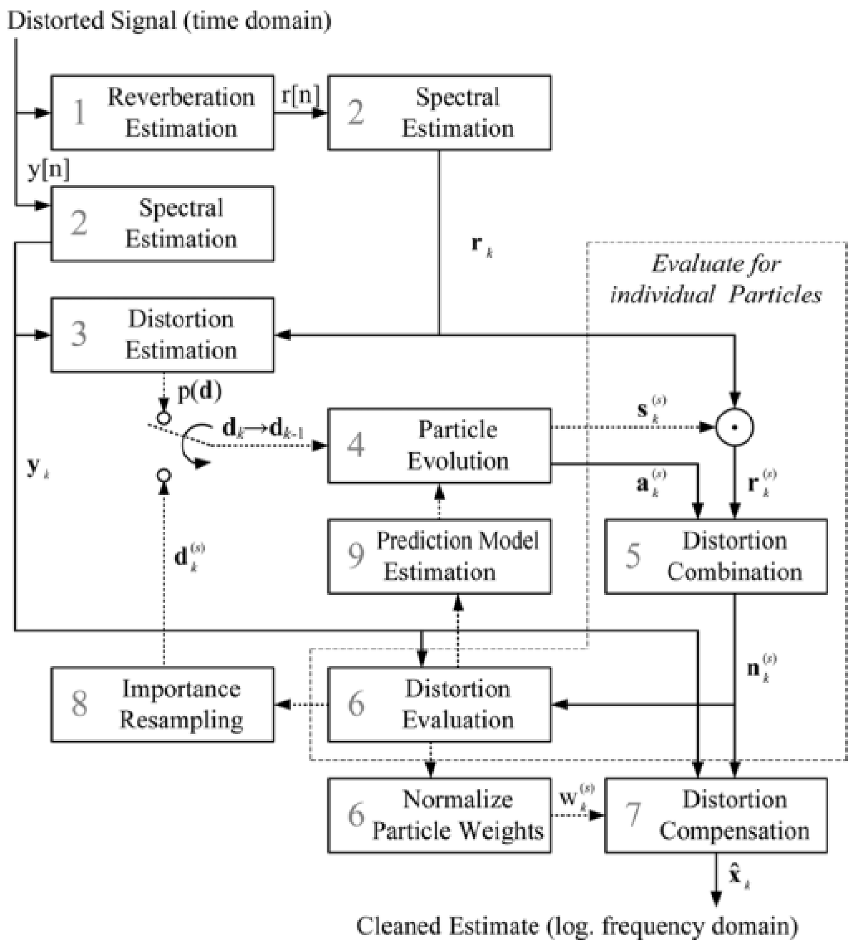
  where $\mu_s = 1.0$ and Σ_s is set to a small variable or can be learned from the data.

Putting The Pieces Together



- Reverberation Estimation
  The reverberation sequence is calculate by MSLP.
- Spectral Estimation
  The reverberation and distorted short-time power spectra are estimated for all frames.
- Distortion Estimation and Particle Initialization
  Samples $d_0^{(s)}, s = 0, \ldots, S-1$ are drawn from the prior distortion density $p(d_0)$
- Particle Evolution
  All particles $d_k^{(s)}, s = 0, \ldots, S-1$ are propagated by the particle transition probability $p(d_k \mid d_{0:k-1})$.

Putting The Pieces Together



- Distortion Combination

  The expected distortion n=u(s,a,r) is calculated as

  $$n[b]_k^{(s)} = \ln\left(e^{a[b]_k^{(s)}} + e^{r[b]_k^{(s)}}\right), \forall b \in B$$

- Distortion Evaluation

  The distortion samples n are evaluated.

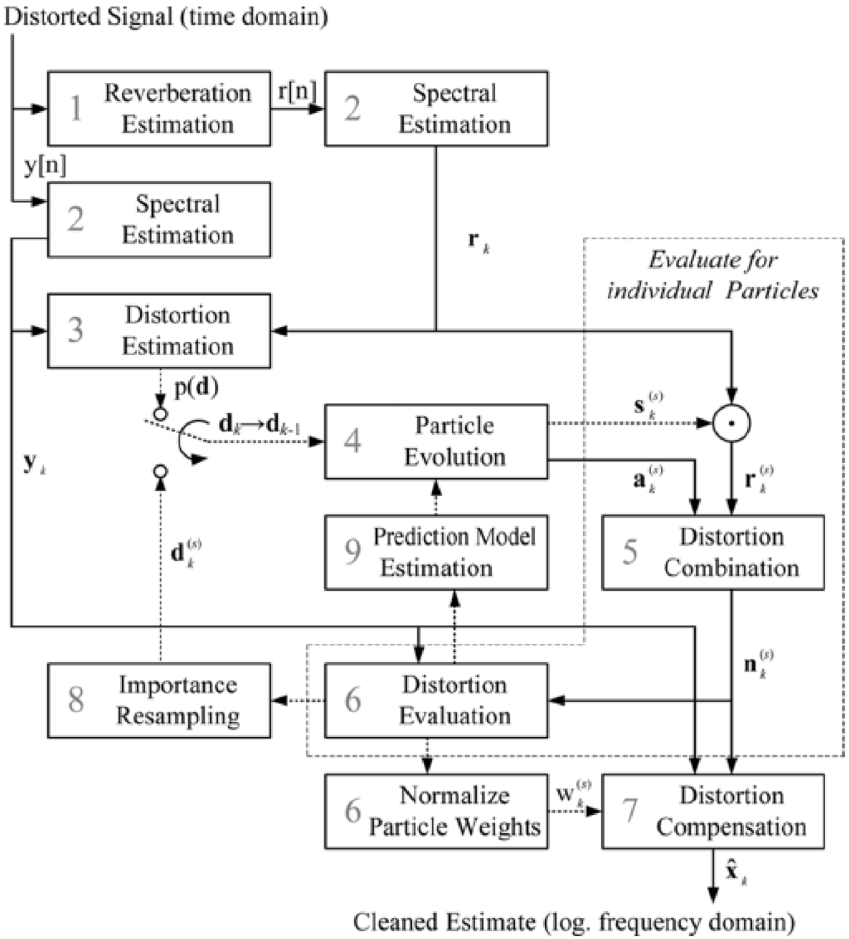
- Distortion Compensation

  The clean feature are calculated.

- Importance Resampling

  Possibly the normalized weights are used to resample among the noise particles $d_k^{(s)}, s = 0, \ldots, S-1$ to prevent the degeneracy problem.

Putting The Pieces Together



- Prediction Model Estimation
  In case of dynamic transition probability model the matrix D_k has to be update.

Step 4 until step 9 are repeated with k->k+1 until either all frames are processed or the track is lost and has to be reinitialized with step 3.

Evaluation Of The Joint Particle Filter Framework

- 35 min lecture speech (continuous, freely spoken) by English with different microphone.
- Janus Recognition Toolkit.
- The optimal step-size D, in MSLP, has been set to 60ms.
- We evaluated on unadpated acoustic models and acoustic models which have been adapted by MLLR and VTLN.

analysis

- SNR: signal-to-noise
- Additive: signal-to-additive-distortion
- Reverberation: signal-to-noise reverberation
- Overall: signal-to-distortions including both distortions

TABLE I

AVERAGE ENERGY OF ADDITIVE NONSTATIONARY DISTORTION AND REVERBERATION VERSUS CLEANED SPEECH ESTIMATE

Microphone	CTM	Lapel	Table Top	Wall
Distance	1 cm	20 cm	150-200 cm	300-400 cm
Estimate	Average Energy vs Cleaned Estimate dB			
SNR	24	23	17	10
Additive	15.1	13.7	12.0	11.3
Reverberation	15.5	11.6	11.5	11.1
Overall	12.3	9.5	8.7	8.2

TABLE II

WORD ERROR RATES FOR DIFFERENT ADDITIVE PARTICLE FILTER ENHANCEMENT TECHNIQUES FOR DIFFERENT SPEAKER TO MICROPHONE DISTANCES

Microphone		CTM		Lapel		Table Top		Wall	
Distance		5 cm		20 cm		150-200 cm		300-400 cm	
Pass		1	2	1	2	1	2	1	2
Prediction	Compensation	Word Error Rate %							
random walk	GMA	11.8	9.4	12.1	9.2	20.9	15.4	49.2	31.6
random walk	SIA	11.6	9.4	12.0	9.2	20.1	15.0	48.6	29.6
static AR	GMA	10.9	9.2	11.3	9.6	18.6	13.7	44.2	26.9
static AR	SIA	10.8	8.9	11.2	9.4	18.5	13.2	42.5	25.3
dynamic AR	GMA	10.8	9.0	11.0	9.2	17.3	13.1	43.5	25.3
dynamic AR	SIA	10.6	9.0	10.7	9.0	17.8	13.2	42.8	25.4

Minimum variance distortionless response (MVDR)

- The MVDR spectrum is a good way of performing all-pole modeling on the speech spectrum.
- Unlike FFT analysis where fixed bandpass filter are used regardless of the characteristics of the incoming signal.
- MVDR obtains the power spectrum estimates by using data-dependent bandpass filters.

Minimum variance distortionless response (MVDR)

- The signal power at a frequency ω_l is determined by filtering the signal by a specially FIR filter $h(n)$ and measuring the power at its output.
- $h(n)$ is designed to minimize its output power subject to the constraint that its response at the frequency of interest ω_l has unity gain :

$$H(e^{j\omega_l}) = \sum_{k=0}^{M} h(k)e^{-j\omega_l k} = 1$$

- This is the *distortionless constraint*

Minimum variance distortionless response (MVDR)

- The distortionless filter $h(n)$ is obtained by solving the following constrained optimization problem

$$\min_{h} \; h^H R_{M+1} h \text{ subject to } v^H(\omega_l) h = 1$$

where $h = [h_0, h_1, \ldots h_M]^H$, $v(\omega) = [1, e^{j\omega}, e^{j2\omega}, \ldots, e^{jM\omega}]$,

R_{M+1} is the $(M+1) \times (M+1)$ Toeplitz autocorrelation matrix of the data

- The solution is $$h_l = \frac{R_{M+1}^{-1} v(\omega_l)}{v^H(\omega_l) R_{M+1}^{-1} v(\omega_l)}$$