GARCH Models and Applications to Speech Enhancement and Anomaly Detection

Prof. Israel Cohen

Electrical Engineering Department
Technion - Israel Institute of Technology

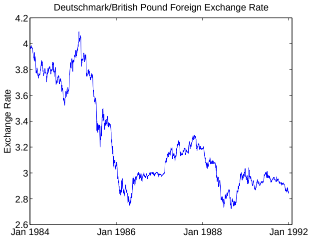
WiSSAP 2016

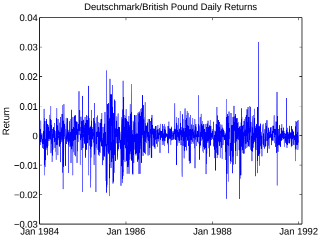
Outline

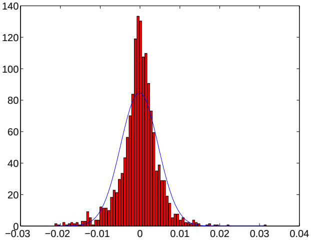
Generalized Autoregressive Conditional Heteroscedasticity (GARCH) Model

- GARCH models [Engle, 1982; Bollerslev, 1986] are widely used in various financial applications such as
  - risk management
  - option pricing
  - foreign exchange.
- They explicitly parameterize the time-varying volatility in terms of past conditional variances and past squared innovations (prediction errors).
- GARCH models take into account excess kurtosis (i.e., heavy tail behavior) and volatility clustering, two important characteristics of financial time-series.

GARCH Model (cont.)







General Form

Let $\{y_t\}$ denote a real-valued discrete-time stochastic process, and let ψ_t denote the information set available at time t.
The innovation process in the MMSE sense is given by

$$\varepsilon_t = y_t - E\{y_t \mid \psi_{t-1}\}$$

and the conditional variance (volatility) of y_t is defined as

$$\sigma_t^2 = \mathrm{var}\{y_t \mid \psi_{t-1}\} = E\left\{\varepsilon_t^2 \mid \psi_{t-1}\right\} .$$

A GARCH model of order (p, q), denoted by $\varepsilon_t \sim \mathrm{GARCH}(p, q)$, has the following general form

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= f\left(\sigma_{t-1}^2, \ldots, \sigma_{t-p}^2, \varepsilon_{t-1}^2 \ldots, \varepsilon_{t-q}^2\right) \end{aligned}$$

where $\{z_t\}$ is a zero-mean unit-variance white noise process with some specified probability distribution.

Linear Formulation

The widely used GARCH model assumes a linear formulation,

$$\sigma_t^2 = \kappa + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \, \sigma_{t-j}^2 \, , \tag{1}$$

and the values of the parameters are constrained by

$$\kappa > 0 \, , \; \alpha_i \geq 0 \, , \; \beta_j \geq 0 \, , \quad i = 1, \ldots, q \, , \; j = 1, \ldots, p \, ,$$

which are sufficient constraints to ensure that the conditional variances $\{\sigma_t^2\}$ are strictly positive. Furthermore, the parameters have to satisfy

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$$

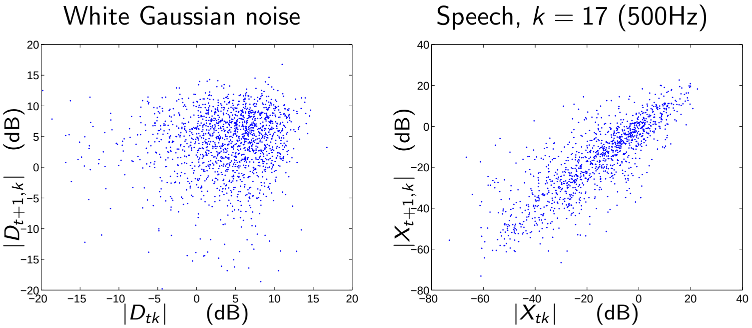
which is a necessary and sufficient constraint for the existence of a finite unconditional variance of the innovations process.

Volatility Clustering and Excess Kurtosis

- GARCH models allow for volatility clustering, since large innovations of either sign increase the variance forecasts for several samples.
- This in return increases the likelihood of large innovations in the succeeding samples, which allows the large innovations to persist.
- Furthermore, the innovations of financial time-series are typically distributed with heavier tails than a Gaussian distribution.

Spectral Analysis

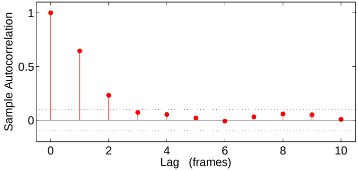
- Clean speech signals, 16 kHz, STFT using Hamming windows, 512 samples length (32 ms), 256 samples framing step (50% overlap).
- Scatter plots for successive spectral magnitudes:



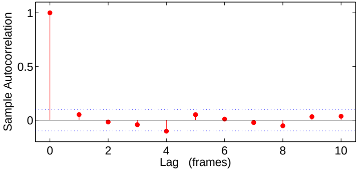
Spectral Analysis (cont.)

- Sample autocorrelation coefficient sequences (ACSs) along time-trajectories:

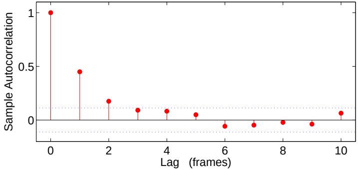
Magnitude, 500Hz, 50% overlap



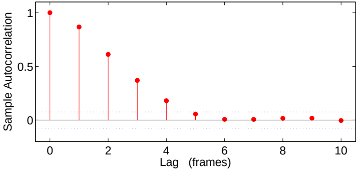
Phase, 500Hz, 50% overlap



Magnitude, 2kHz, 50% overlap

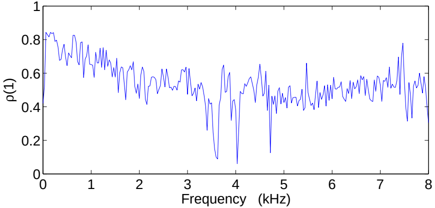


Magnitude, 500Hz, 75% overlap

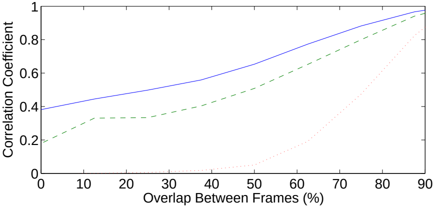


Spectral Analysis (cont.)

- Typical variation of $\rho(1)$, the correlation coefficient between successive spectral magnitudes:



Variation on frequency



Variation on overlap
1kHz (solid)
2kHz (dashed)
noise (dotted)

Spectral Analysis (cont.)

- When observing a time series of successive expansion coefficients in a fixed frequency bin, successive magnitudes of the expansion coefficients are highly correlated, whereas successive phases are nearly uncorrelated.
- Hence, the expansion coefficients are clustered in the sense that large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes, while the phase is unpredictable.

Speech signals in the STFT domain are characterized by volatility clustering and heavy-tailed distribution.

Problem Formulation

Let $\{Y_{tk}\}$ denote a noisy speech signal in the STFT domain:

$$\begin{aligned} H_1^{tk} \text{ (speech present)}: \quad Y_{tk} &= X_{tk} + D_{tk} \\ H_0^{tk} \text{ (speech absent)}: \quad Y_{tk} &= D_{tk} \,. \end{aligned}$$

The spectral enhancement problem can be formulated as

$$\min_{\hat{X}_{tk}} E\left\{ d\left(X_{tk}, \hat{X}_{tk}\right) \;\middle|\; \hat{p}_{tk}\,, \hat{\lambda}_{tk}\,, \widehat{\sigma^2_{tk}}\,,\; Y_{tk} \right\}$$

- $d\left(X_{tk}, \hat{X}_{tk}\right)$ - distortion measure between X_{tk} and $\hat{X}_{tk}$
- $\hat{p}_{tk} = P\left(H_1^{tk} \mid \psi_t\right)$ - speech presence probability estimate
- $\hat{\lambda}_{tk} = E\left\{|X_{tk}|^2 \mid H_1^{tk}, \psi_t\right\}$ - speech spectral variance estimate
- $\widehat{\sigma^2_{tk}} = E\left\{|Y_{tk}|^2 \mid H_0^{tk}, \psi_t\right\}$ - noise spectral variance estimate
- ψ_t - information employed for estimation at frame t

Problem Formulation (cont.)

In particular, assuming a squared error distortion measure of the form

$$d\left(X_{tk}, \hat{X}_{tk}\right) = \left|g(\hat{X}_{tk}) - \tilde{g}(X_{tk})\right|^2$$

where $g(X)$ and $\tilde{g}(X)$ are specific functions of X (*e.g.*, X, $|X|$, $\log|X|$, $e^{j\angle X}$)

the estimator $\hat{X}_{tk}$ is calculated from

$$\begin{aligned} g(\hat{X}_{tk}) &= E\left\{\tilde{g}(X_{tk}) \,\middle|\, \hat{p}_{tk}\,, \hat{\lambda}_{tk}\,, \widehat{\sigma^2_{tk}}\,,\, Y_{tk}\right\} \\ &= \hat{p}_{tk}\, E\left\{\tilde{g}(X_{tk}) \,\middle|\, H_1^{tk}\,, \hat{\lambda}_{tk}\,, \widehat{\sigma^2_{tk}}\,,\, Y_{tk}\right\} \\ &\quad +(1-\hat{p}_{tk})\, E\left\{\tilde{g}(X_{tk}) \,\middle|\, H_0^{tk}\,,\, Y_{tk}\right\}. \end{aligned}$$

Problem Formulation (cont.)

The design of a particular estimator for X_{tk} requires the following specifications:

- Functions $g(X)$ and $\tilde{g}(X)$, which determine the fidelity criterion of the estimator.
- A conditional probability density function (pdf) $p\left(X_{tk} \mid \lambda_{tk}\,,\, H_1^{tk}\right)$ for X_{tk} under H_1^{tk} given its variance λ_{tk}, which determines the statistical model.
- An estimator $\hat{\lambda}_{tk}$ for the speech spectral variance.
- An estimator $\widehat{\sigma_{tk}^2}$ for the noise spectral variance.
- An estimator $\hat{p}_{tk|t-1} = P\left(H_1^{tk} \mid \psi_{t-1}\right)$ for the *a priori* speech presence probability, where ψ_{t-1} represents the information set known prior to having the measurement Y_{tk}.

GARCH Modeling

- Given $\{\lambda_{tk}\}$ and the state of speech presence in each time-frequency bin (H_1^{tk} or H_0^{tk}), the speech spectral coefficients $\{X_{tk}\}$ are generated by

$$X_{tk} = \sqrt{\lambda_{tk}}\, V_{tk}$$

where $\left\{ V_{tk} \,|\, H_0^{tk} \right\}$ are identically zero, and $\left\{ V_{tk} \,|\, H_1^{tk} \right\}$ are statistically independent complex random variables with zero mean, unit variance, and iid real and imaginary parts:

$$\begin{array}{ll} H_1^{tk}: & E\left\{V_{tk}\right\} = 0\,,\ E\left\{|V_{tk}|^2\right\} = 1 \\ H_0^{tk}: & V_{tk} = 0 \end{array}$$

- The speech spectral variances $\{\lambda_{tk}\}$ are hidden from direct observation even under perfect conditions of zero noise ($D_{tk} = 0$ for all tk).

GARCH Modeling (cont.)

- Over the past decades, the decision-directed approach has become the acceptable estimation method for variances of speech spectral coefficients [Ephraim and Malah, 1984]

$$\hat{\lambda}_{tk} = \max\left\{\alpha \left|\hat{X}_{t-1,k}\right|^2 + (1-\alpha)\left(\left|Y_{tk}\right|^2 - \sigma_{tk}^2\right) \, , \, \xi_{\min}\, \sigma_{tk}^2\right\} .$$

GARCH Modeling (cont.)

- Over the past decades, the decision-directed approach has become the acceptable estimation method for variances of speech spectral coefficients [Ephraim and Malah, 1984]

$$\hat{\lambda}_{tk} = \max \left\{ \alpha \left| \hat{X}_{t-1,k} \right|^2 + (1-\alpha) \left(|Y_{tk}|^2 - \sigma_{tk}^2 \right) \, , \, \xi_{\min} \, \sigma_{tk}^2 \right\} \, .$$

- The decision-directed approach is not supported by a statistical model.
- α and $\xi_{\min}$ have to be determined by simulations and subjective listening tests for each particular setup of time-frequency transformation and speech enhancement algorithm.
- α and $\xi_{\min}$ are not adapted to the speech components.

GARCH Modeling (cont.)

- Our approach is to assume that $\{\lambda_{tk}\}$ are random variables, and to introduce *conditional* variances which are estimated from the available information.
- Let $\lambda_{tk|\tau} \triangleq E\left\{|X_{tk}|^2 \mid H_1^{tk}\,,\, \mathcal{X}_0^{\tau}\right\}$ denote the *conditional* variance of X_{tk} under H_1^{tk} given the clean spectral coefficients up to frame τ. We assume that $\lambda_{tk|t-1}$, referred to as the *one-frame-ahead conditional variance*, is a random process which evolves as a $\mathrm{GARCH}(1,1)$ process:

$$\lambda_{tk|t-1} = \lambda_{\min} + \mu\,|X_{t-1,k}|^2 + \delta\left(\lambda_{t-1,k|t-2} - \lambda_{\min}\right)$$

where

$$\lambda_{\min} > 0\,, \quad \mu \geq 0\,, \quad \delta \geq 0\,, \quad \mu + \delta < 1$$

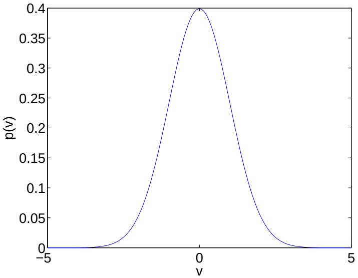
are the standard constraints imposed on the parameters of the GARCH model.

GARCH Modeling (cont.)

- Gaussian model

$$p\left(V_{\rho tk} \mid H_1^{tk}\right) = \frac{1}{\sqrt{\pi}} \exp\left(-V_{\rho tk}^2\right)$$

$$\rho \in \{R, I\}\ , V_{Rtk} \triangleq \Re\{V_{tk}\}\ , V_{Itk} \triangleq \Im\{V_{tk}\}$$



GARCH Modeling (cont.)

- Gaussian model

$$p\left(V_{\rho tk} \mid H_1^{tk}\right) = \frac{1}{\sqrt{\pi}} \exp\left(-V_{\rho tk}^2\right)$$

$\rho \in \{R, I\}$, $V_{Rtk} \triangleq \Re\{V_{tk}\}$, $V_{Itk} \triangleq \Im\{V_{tk}\}$

- Gamma model

$$p\left(V_{\rho tk} \mid H_1^{tk}\right) = \frac{1}{2\sqrt{\pi}} \left(\frac{3}{2}\right)^{1/4} |V_{\rho tk}|^{-1/2} \exp\left(-\sqrt{\frac{3}{2}}\,|V_{\rho tk}|\right)$$

- Laplacian model

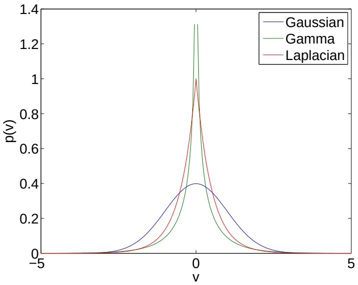
$$p\left(V_{\rho tk} \mid H_1^{tk}\right) = \exp\left(-2\,|V_{\rho tk}|\right) .$$

GARCH Modeling (cont.)

- Gaussian model

$$p\left(V_{\rho tk} \mid H_1^{tk} \right) = \frac{1}{\sqrt{\pi}} \exp\left(-V_{\rho tk}^2 \right)$$

$$\rho \in \{R, I\} \;, V_{Rtk} \triangleq \Re\{V_{tk}\} \;, V_{Itk} \triangleq \Im\{V_{tk}\}$$



Variance Estimation

Following the rational of Kalman filtering:

- Start with an estimate $\hat{\lambda}_{tk|t-1}$, and update the variance by using the additional information Y_{tk},

Update step:

$$\hat{\lambda}_{tk|t} = E\left\{ |X_{tk}|^2 \,\middle|\, \hat{\lambda}_{tk|t-1}\,,\, Y_{tk} \right\}$$

- Propagate the variance estimate ahead in time to obtain a conditional variance estimate at frame $t+1$,

Propagation step:

$$\hat{\lambda}_{t+1,k|t} = \lambda_{\min} + \mu\,\hat{\lambda}_{tk|t} + \delta\left(\hat{\lambda}_{tk|t-1} - \lambda_{\min}\right)$$

- The propagation and update steps are iterated, to recursively estimate the speech variances as new data arrive.

Relation to Decision-Directed Estimation

- Recall the *heuristically motivated* decision-directed estimator [Ephraim and Malah, 1984]

$$\hat{\lambda}_{tk} = \max\left\{\alpha\,|\hat{X}_{t-1,k}|^2 + (1-\alpha)\left(|Y_{tk}|^2 - \sigma_{tk}^2\right)\ ,\ \xi_{\min}\,\sigma_{tk}^2\right\}$$

- A special case of the GARCH-based variance estimator degenerates to a decision-directed estimator with a *time-varying frequency-dependent* weighting factor α_{tk}

$$\alpha \iff \alpha_{tk}$$

$$\xi_{\min}\,\sigma_{tk}^2 \iff \lambda_{\min}$$

$$\left|\hat{X}_{t-1,k}\right|^2 \iff \hat{\lambda}_{t-1,k|t-1} \triangleq E\left\{|X_{t-1,k}|^2 \,\middle|\, \hat{\lambda}_{t-1,k|t-2}\,,\, Y_{t-1,k}\right\}$$

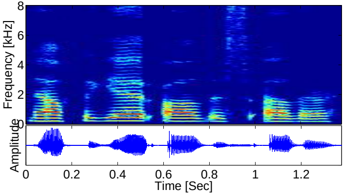
Experimental Results

- Speech signals: 20 different utterances from 20 different speakers, sampled at 16 kHz and degraded by white Gaussian noise with SNRs in the range $[0, 20]$dB.
- Eight different speech enhancement algorithms are compared

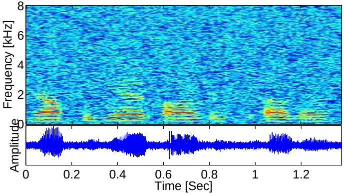
Algorithm #	Statistical Model	Variance Estimation	Fidelity Criterion
1	Gaussian	GARCH	MMSE
2	Gamma	GARCH	MMSE
3	Laplacian	GARCH	MMSE
4	Gaussian	Decision-Directed	MMSE
5	Gamma	Decision-Directed	MMSE
6	Laplacian	Decision-Directed	MMSE
7	Gaussian	GARCH	MMSE-LSA
8	Gaussian	Decision-Directed	MMSE-LSA

Experimental Results (cont.)

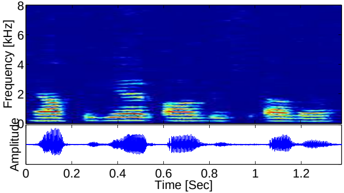
Clean speech signal



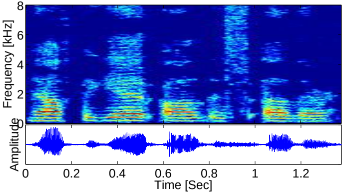
Noisy signal, SNR $= 5$dB
LSD $= 13.75$dB, PESQ$= 1.76$



Decision-Directed, MMSE-LSA
LSD $= 9.00$dB, PESQ$= 2.57$



GARCH, MMSE-LSA
LSD $= 3.59$dB, PESQ$= 2.88$



Experimental Results (cont.)

Log-Spectral Distortion (LSD)

Input SNR [dB]	GARCH modeling method Gaussian MMSE	GARCH Gaussian LSA	GARCH Gamma MMSE	GARCH Laplacian MMSE	Decision-Directed method Gaussian MMSE	DD Gaussian LSA	DD Gamma MMSE	DD Laplacian MMSE
0	7.77	**4.85**	8.03	7.91	18.89	11.35	17.76	18.14
5	5.78	**4.04**	6.93	6.45	17.29	11.03	15.73	16.26
10	4.14	**3.27**	5.35	4.85	13.87	9.13	11.83	12.48
15	2.50	**2.25**	3.23	2.92	9.19	6.05	6.95	7.59
20	1.30	**1.28**	1.55	1.44	4.88	3.13	2.88	3.34

Perceptual Evaluation of Speech Quality (PESQ) scores (ITU-T P.862)

Input SNR [dB]	GARCH modeling method Gaussian MMSE	GARCH Gaussian LSA	GARCH Gamma MMSE	GARCH Laplacian MMSE	Decision-Directed method Gaussian MMSE	DD Gaussian LSA	DD Gamma MMSE	DD Laplacian MMSE
0	2.52	**2.55**	2.47	2.48	1.91	2.21	1.98	1.96
5	2.97	**2.98**	2.90	2.91	2.30	2.61	2.38	2.36
10	3.37	**3.38**	3.28	3.31	2.70	2.99	2.77	2.75
15	3.67	**3.69**	3.59	3.62	3.09	3.31	3.17	3.15
20	3.88	**3.89**	3.83	3.85	3.53	3.64	3.62	3.60

Experimental Results (cont.)

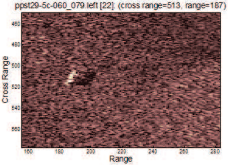
- The GARCH modeling method yields lower LSD and higher PESQ scores than the decision-directed method.
- Using the decision-directed method, a Gaussian model is inferior to Gamma and Laplacian models.
- Using the GARCH modeling method, a Gaussian model is superior to Gamma and Laplacian models.
- It is difficult, or even impossible, to derive analytical expressions for MMSE-LSA estimators under Gamma or Laplacian models.

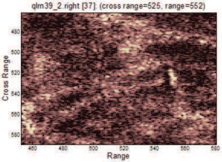
The GARCH modeling method facilitates MMSE-LSA estimation, while taking into consideration the heavy-tailed distribution.

Anomaly Detection

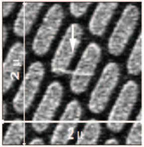
- Anomaly detection approach is attractive when target models are not available or are unreliable.

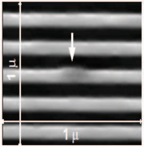
Sea mines

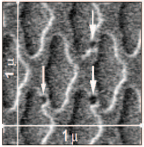




Wafer defects

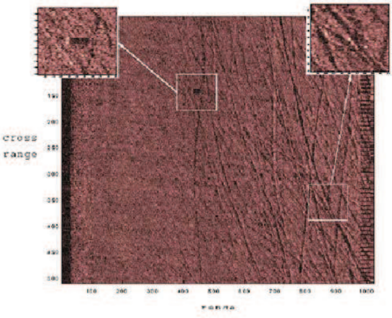






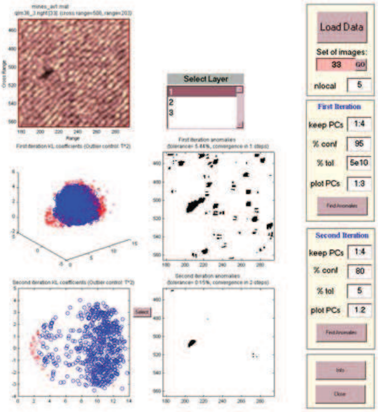
Sea-Mine Detection

- Mine detection in sonar imagery is a challenging problem due to the large variability of background clutter and the object characteristics.

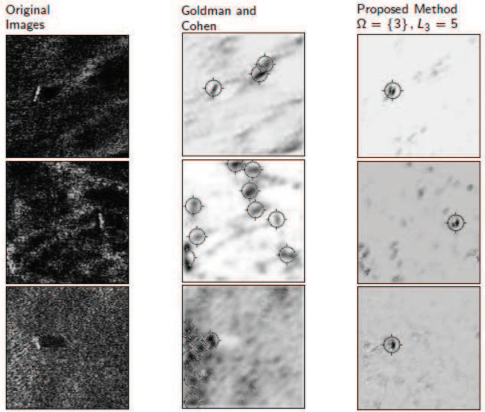


Sea-Mine Detection (cont.)

- The variability of the target signature is described using a subspace model.
- The variability of the background is described using a multidimensional GARCH model .
- The GARCH model characterizes the heavy tails and clustering of innovations in the background.



Experimental Results



Conclusions

- GARCH modeling provides a new framework for speech enhancement and anomaly detection in adverse environments.
- GARCH models take into account excess kurtosis and volatility clustering, two important characteristics of financial time-series, speech signals, and background clutter in sonar imagery.
- The decision-directed approach, which is heuristically motivated, can be obtained as a special case of GARCH-based variance estimation.
- GARCH modeling enables MMSE log-spectral amplitude estimation of speech while taking into consideration the heavy-tailed distribution.