New Statistical Model for the Enhancement of Noisy Speech

Prof. Israel Cohen

Electrical Engineering Department
Technion - Israel Institute of Technology

February 22, 2007
Outline

1. Problem Formulation and Motivation
   - Spectral Enhancement
   - Speech Models
   - Objectives

2. New Speech Models
   - Spectral Analysis
   - GARCH Model
   - Speech Modeling

3. Speech Enhancement
   - Variance Estimation
   - Relation to Decision-Directed Estimation
   - Experimental Results

4. Conclusions and Future Directions
Spectral Enhancement

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

\[
H_{1}^{tk} \text{ (speech present)} : \quad Y_{tk} = X_{tk} + D_{tk}
\]

\[
H_{0}^{tk} \text{ (speech absent)} : \quad Y_{tk} = D_{tk}.
\]

The spectral enhancement problem can be formulated as

\[
\min_{\hat{X}_{tk}} \mathbb{E} \left\{ d \left( X_{tk}, \hat{X}_{tk} \right) \bigg| \hat{p}_{tk}, \hat{\lambda}_{tk}, \hat{\sigma}_{tk}^{2}, Y_{tk} \right\}
\]

- \( d \left( X_{tk}, \hat{X}_{tk} \right) \) - distortion measure between \( X_{tk} \) and \( \hat{X}_{tk} \)
- \( \hat{p}_{tk} = P \left( H_{1}^{tk} \big| \psi_{t} \right) \) - speech presence probability estimate
- \( \hat{\lambda}_{tk} = \mathbb{E} \left\{ |X_{tk}|^2 \big| H_{1}^{tk}, \psi_{t} \right\} \) - speech spectral variance estimate
- \( \hat{\sigma}_{tk}^{2} = \mathbb{E} \left\{ |Y_{tk}|^2 \big| H_{0}^{tk}, \psi_{t} \right\} \) - noise spectral variance estimate
- \( \psi_{t} \) - information employed for estimation at frame \( t \) (e.g., noisy data observed through time \( t \))
Spectral Enhancement (cont.)

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

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

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

\[
g(\hat{X}_{tk}) = E \left\{ \tilde{g}(X_{tk}) \left| \hat{p}_{tk}, \hat{\lambda}_{tk}, \hat{\sigma}_{tk}^2, Y_{tk} \right. \right\} \\
= \hat{p}_{tk} E \left\{ \tilde{g}(X_{tk}) \left| H_{1}^{tk}, \hat{\lambda}_{tk}, \hat{\sigma}_{tk}^2, Y_{tk} \right. \right\} \\
+ (1 - \hat{p}_{tk}) E \left\{ \tilde{g}(X_{tk}) \left| H_{0}^{tk}, Y_{tk} \right. \right\}.
\]
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) $p\left(X_{tk} \mid \lambda_{tk}, H_{tk}^1\right)$ for $X_{tk}$ under $H_{tk}^1$ given its variance $\lambda_{tk}$, which determines the statistical model.
- An estimator $\hat{\lambda}_{tk}$ for the speech spectral variance.
- An estimator $\hat{\sigma}_{tk}^2$ for the noise spectral variance.
- An estimator $\hat{p}_{tk|t-1} = P\left(H_{tk}^1 \mid \psi_{t-1}\right)$ for the a priori speech presence probability, where $\psi_{t-1}$ represents the information set known prior to having the measurement $Y_{tk}$. 
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 \( \{ V_{tk} \mid H_{0}^{tk} \} \) are identically zero, and \( \{ V_{tk} \mid H_{1}^{tk} \} \) are statistically independent complex random variables with zero mean, unit variance, and iid real and imaginary parts:

\[
H_{1}^{tk} : \quad E \{ V_{tk} \} = 0 , \quad E \{ |V_{tk}|^2 \} = 1
\]

\[
H_{0}^{tk} : \quad V_{tk} = 0
\]
Speech Models (cont.)

- Gaussian model [McAulay and Malpass, 1980; Ephraim and Malah, 1984]

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

\[ \rho \in \{ R, I \} , \quad V_{Rtk} \triangleq \Re \{ V_{tk} \} , \quad V_{I tk} \triangleq \Im \{ V_{tk} \} \]
Speech Models (cont.)

- **Gaussian model** [McAulay and Malpass, 1980; Ephraim and Malah, 1984]

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

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

- **Gamma model** [Porter and Boll, 1984; Martin, 2002]

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

- **Laplacian model** [Martin and Breithaupt, 2003; Lotter and Vary, 2003]

  \[
p \left( V_{\rho tk} \mid H_{1}^{tk} \right) = \exp \left( -2 \left| V_{\rho tk} \right| \right).
  \]
Gaussian model [McAulay and Malpass, 1980; Ephraim and Malah, 1984]

\[ p \left( V_{\rho tk} \mid H_{1tk}^t \right) = \frac{1}{\sqrt{\pi}} \exp \left( -V_{\rho tk}^2 \right) \]

\[ \rho \in \{ R, I \} \ , \ V_{Rtk} \triangleq \Re \{ V_{tk} \} \ , \ V_{Itk} \triangleq \Im \{ V_{tk} \} \]
Over the past two decades, the decision-directed approach has become the acceptable estimation method for variances of speech spectral coefficients [Ephraim and Malah, 1984]

\[
\hat{\lambda}_{tk} = \max \left\{ \alpha |\hat{X}_{t-1,k}|^2 + (1 - \alpha) \left( |Y_{tk}|^2 - \sigma_{tk}^2 \right) , \xi \min \sigma_{tk}^2 \right\}.
\]
Over the past two decades, the decision-directed approach has become the acceptable estimation method for variances of speech spectral coefficients [Ephraim and Malah, 1984]

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

The decision-directed approach is not supported by a statistical model.

\( \alpha \) 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.

\( \alpha \) and \( \xi_{\min} \) are not adapted to the speech components.
Objectives

- Analyze the time-frequency correlation of speech and noise signals in the STFT domain.
- Formulate statistical models for speech signals in the STFT domain, which take into consideration the time-frequency correlation and heavy-tailed distribution of the expansion coefficients.
- Derive estimators for the speech spectral variances, which are based on the proposed models.
- Show that a special case of the proposed variance estimator degenerates to a “decision-directed” estimator.
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:
  - White Gaussian noise
  - Speech, $k = 17$ (500Hz)
Sample autocorrelation coefficient sequences (ACSs) along time-trajectories:

**Magnitude, 500Hz, 50% overlap**

**Phase, 500Hz, 50% overlap**

**Magnitude, 2kHz, 50% overlap**

**Magnitude, 500Hz, 75% overlap**
Typical variation of $\rho(1)$, the correlation coefficient between successive spectral magnitudes:

Variation on frequency

Variation on overlap

1kHz (solid)
2kHz (dashed)
noise (dotted)
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.
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, and foreign exchange.
- They explicitly parameterize the time-varying volatility in terms of past conditional variances and past squared innovations (prediction errors), while taking into account excess kurtosis (i.e., heavy tail behavior) and volatility clustering, two important characteristics of financial time-series.

⇒ Modeling speech expansion coefficients as GARCH processes offers a reasonable model on which to base the variance estimation, while taking into consideration the heavy-tailed distribution.
GARCH Model (cont.)

Deutschmark/British Pound Foreign Exchange Rate

Deutschmark/British Pound Daily Returns

Prof. Israel Cohen
Technion - Israel Institute on Technology
Let \( \{y_t\} \) denote a real-valued discrete-time stochastic process, and let \( \psi_t \) 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

\[
\sigma_t^2 = \text{var} \{ y_t \mid \psi_{t-1} \} = E \{ \varepsilon_t^2 \mid \psi_{t-1} \}.
\]

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

\[
\begin{align*}
\varepsilon_t &= \sigma_t z_t \\
\sigma_t^2 &= f(\sigma_{t-1}^2, \ldots, \sigma_{t-p}^2, \varepsilon_{t-1}^2 \ldots, \varepsilon_{t-q}^2)
\end{align*}
\]

where \( \{z_t\} \) is a zero-mean unit-variance white noise process with some specified probability distribution.
The widely used GARCH model assumes a linear formulation,

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

(1)

and the values of the parameters are constrained by

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

(sufficient constraints to ensure that the conditional variances \( \{\sigma_t^2\} \) are strictly positive) and by

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

(necessary and sufficient constraint for the existence of a finite unconditional variance of the innovations process).
As before, given \( \{ \lambda_{tk} \} \) and the state of speech presence in each time-frequency bin (\( H^t_{1k} \) or \( H^t_{0k} \)), \( \{ X_{tk} \} \) are generated by

\[
X_{tk} = \sqrt{\lambda_{tk}} \ V_{tk}
\]

where \( \{ V_{tk} \mid H^t_{1k} \} \) are statistically independent complex random variables

\[
H^t_{1k} : \quad E\{ V_{tk} \} = 0, \quad E\{ |V_{tk}|^2 \} = 1
\]
\[
H^t_{0k} : \quad V_{tk} = 0
\]

However, \( \{ \lambda_{tk} \} \) are hidden from direct observation even under perfect conditions of zero noise (\( D_{tk} = 0 \) for all \( tk \)).
Our approach is to assume that \( \{\lambda_{tk}\} \) themselves are random variables, and to introduce *conditional* variances which are estimated from the available information.

Let \( \lambda_{tk|\tau} \triangleq E \left\{ |X_{tk}|^2 \mid H_{1}^{tk}, X_{0}^{\tau} \right\} \) denote the *conditional* variance of \( X_{tk} \) under \( H_{1}^{tk} \) given the clean spectral coefficients up to frame \( \tau \). We assume that \( \lambda_{tk|t-1} \), referred to as the *one-frame-ahead conditional variance*, is a random process which evolves as a GARCH(1, 1) process:

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

where

\[
\lambda_{\text{min}} > 0, \quad \mu \geq 0, \quad \delta \geq 0, \quad \mu + \delta < 1
\]

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

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

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

**Propagation step:**

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

- The propagation and update steps are iterated, to recursively estimate the speech variances as new data arrive.
For a Gaussian-GARCH model, the update step can be written as

\[ \hat{\lambda}_{tk|t} = \alpha_{tk} \hat{\lambda}_{tk|t-1} + (1 - \alpha_{tk}) \left( |Y_{tk}|^2 - \sigma_{tk}^2 \right) \]

with

\[ \alpha_{tk} \triangleq 1 - \frac{\hat{\lambda}_{tk|t-1}^2}{\left( \hat{\lambda}_{tk|t-1} + \sigma_{tk}^2 \right)^2} \cdot \]

Using the propagation step with \( \mu \equiv 1 \) and applying the lower bound constraint to \( \hat{\lambda}_{tk|t} \) rather than \( \hat{\lambda}_{tk|t-1} \), we have

\[ \hat{\lambda}_{tk|t} = \max \left\{ \alpha_{tk} \hat{\lambda}_{t-1,k|t-1} + (1 - \alpha_{tk}) \left( |Y_{tk}|^2 - \sigma_{tk}^2 \right), \lambda_{\text{min}} \right\} \]
Recall the *heuristically motivated* decision-directed estimator

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

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

\[ \begin{align*}
\alpha & \iff \alpha_{tk} \\
\xi_{\min} \sigma_{tk}^2 & \iff \lambda_{\min} \\
|\hat{X}_{t-1,k}|^2 & \iff \hat{\lambda}_{t-1,k|t-1} \Delta E \left\{ |X_{t-1,k}|^2 \left| \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k} \right. \right\}
\end{align*} \]
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

<table>
<thead>
<tr>
<th>Algorithm #</th>
<th>Statistical Model</th>
<th>Variance Estimation</th>
<th>Fidelity Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gaussian</td>
<td>GARCH</td>
<td>MMSE</td>
</tr>
<tr>
<td>2</td>
<td>Gamma</td>
<td>GARCH</td>
<td>MMSE</td>
</tr>
<tr>
<td>3</td>
<td>Laplacian</td>
<td>GARCH</td>
<td>MMSE</td>
</tr>
<tr>
<td>4</td>
<td>Gaussian</td>
<td>Decision-Directed</td>
<td>MMSE</td>
</tr>
<tr>
<td>5</td>
<td>Gamma</td>
<td>Decision-Directed</td>
<td>MMSE</td>
</tr>
<tr>
<td>6</td>
<td>Laplacian</td>
<td>Decision-Directed</td>
<td>MMSE</td>
</tr>
<tr>
<td>7</td>
<td>Gaussian</td>
<td>GARCH</td>
<td>MMSE-LSA</td>
</tr>
<tr>
<td>8</td>
<td>Gaussian</td>
<td>Decision-Directed</td>
<td>MMSE-LSA</td>
</tr>
</tbody>
</table>
Experimental Results (cont.)

Clean speech signal

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

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

GARCH, MMSE-LSA
LSD = 3.59dB, PESQ = 2.88
## Experimental Results (cont.)

### Log-Spectral Distortion (LSD)

<table>
<thead>
<tr>
<th>Input SNR [dB]</th>
<th>GARCH modeling method</th>
<th>Decision-Directed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gaussian MMSE</td>
<td>Gamma MMSE</td>
</tr>
<tr>
<td>0</td>
<td>7.77</td>
<td>4.85</td>
</tr>
<tr>
<td>5</td>
<td>5.78</td>
<td>4.04</td>
</tr>
<tr>
<td>10</td>
<td>4.14</td>
<td>3.27</td>
</tr>
<tr>
<td>15</td>
<td>2.50</td>
<td>2.25</td>
</tr>
<tr>
<td>20</td>
<td>1.30</td>
<td>1.28</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Input SNR [dB]</th>
<th>GARCH modeling method</th>
<th>Decision-Directed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gaussian MMSE</td>
<td>LSA MMSE</td>
</tr>
<tr>
<td>0</td>
<td>2.52</td>
<td>2.55</td>
</tr>
<tr>
<td>5</td>
<td>2.97</td>
<td>2.98</td>
</tr>
<tr>
<td>10</td>
<td>3.37</td>
<td>3.38</td>
</tr>
<tr>
<td>15</td>
<td>3.67</td>
<td>3.69</td>
</tr>
<tr>
<td>20</td>
<td>3.88</td>
<td>3.89</td>
</tr>
</tbody>
</table>
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.
Conclusions and Future Directions

- The decision-directed approach is heuristically motivated, and cannot be adapted to the components of the speech signal.
- Speech signals in the STFT domain demonstrate both “volatility clustering” and heavy tail behavior.
- GARCH modeling provides a new framework for optimal restoration of speech signals in adverse environments.
- The concepts of GARCH and hidden Markov models can be used to develop Markov-Switching GARCH models for multichannel speech processing, voice activity detection, source localization, dereverberation and adaptive beamforming in noisy and reverberant environments.