

# Speech Enhancement






Akihiro Kato  
University of Eastern Finland



# Overview

- Noisy speech
  - Time-domain / Complex frequency domain
  - Magnitude and phase spectra
- Signal-to-noise ratios
- Filtering-Based Speech Enhancement
  - Spectral subtraction
  - Noise estimation
  - Other methods of filtering-based speech enhancement
- Model-based speech enhancement

# Noisy Speech

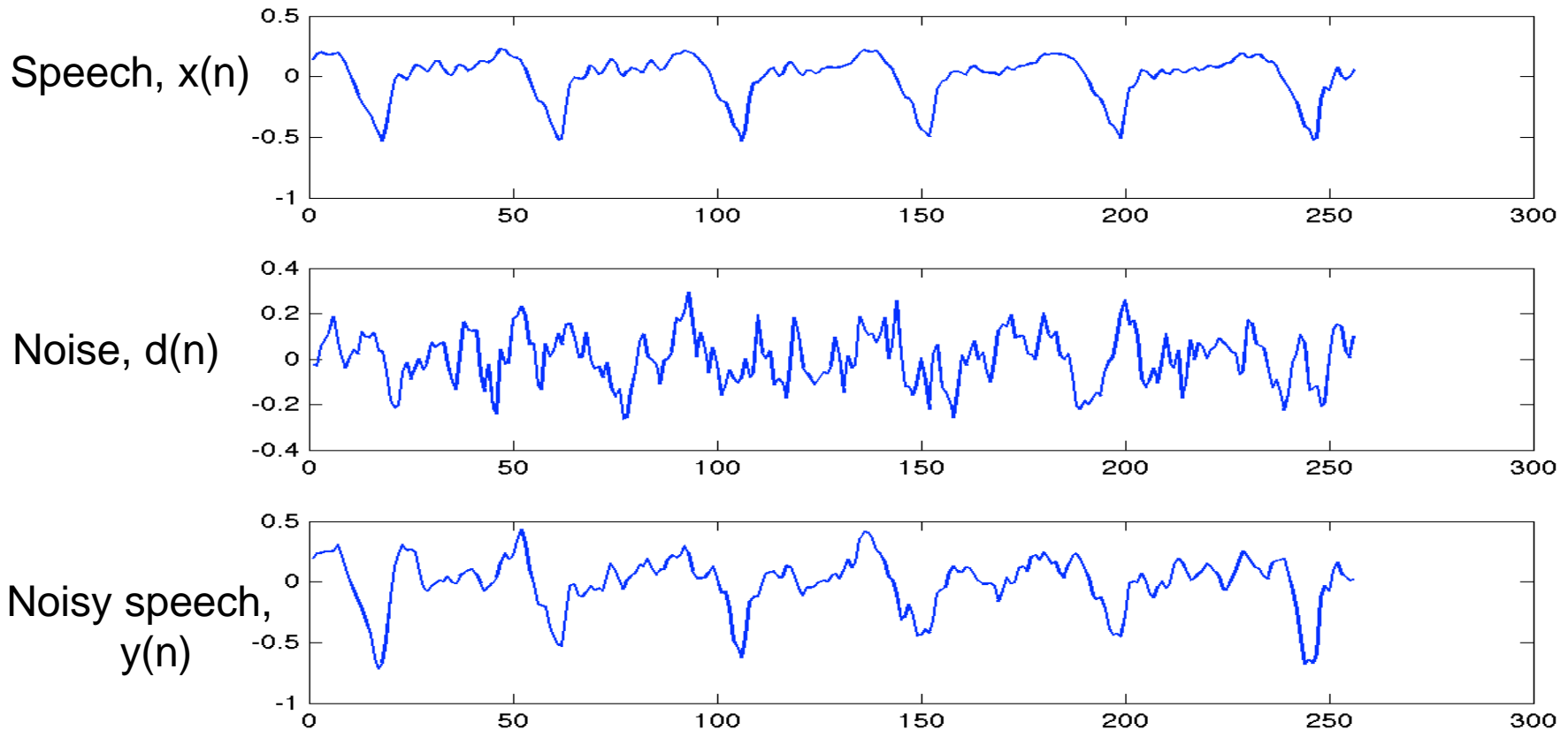
- In real applications of speech processing technology many artefacts can distort or corrupt the speech
- Acoustic noise 
  - Many sources – traffic, people, etc
- Radio noise 
- Packet loss/frame errors 
  - Common on VoIP and mobile telephony
- Channel distortion 
  - Quality of handset varies
- Echo 
  - Between earpiece and microphone

# Noisy Speech – Time-Domain

- Given a time domain speech signal,  $x(n)$ , a noise signal,  $d(n)$ , can be added to create a noisy speech signal,  $y(n)$

$$y(n) = x(n) + d(n)$$

- Assuming no non-linearities (e.g. microphone or amplifier clipping) then the operation is additive

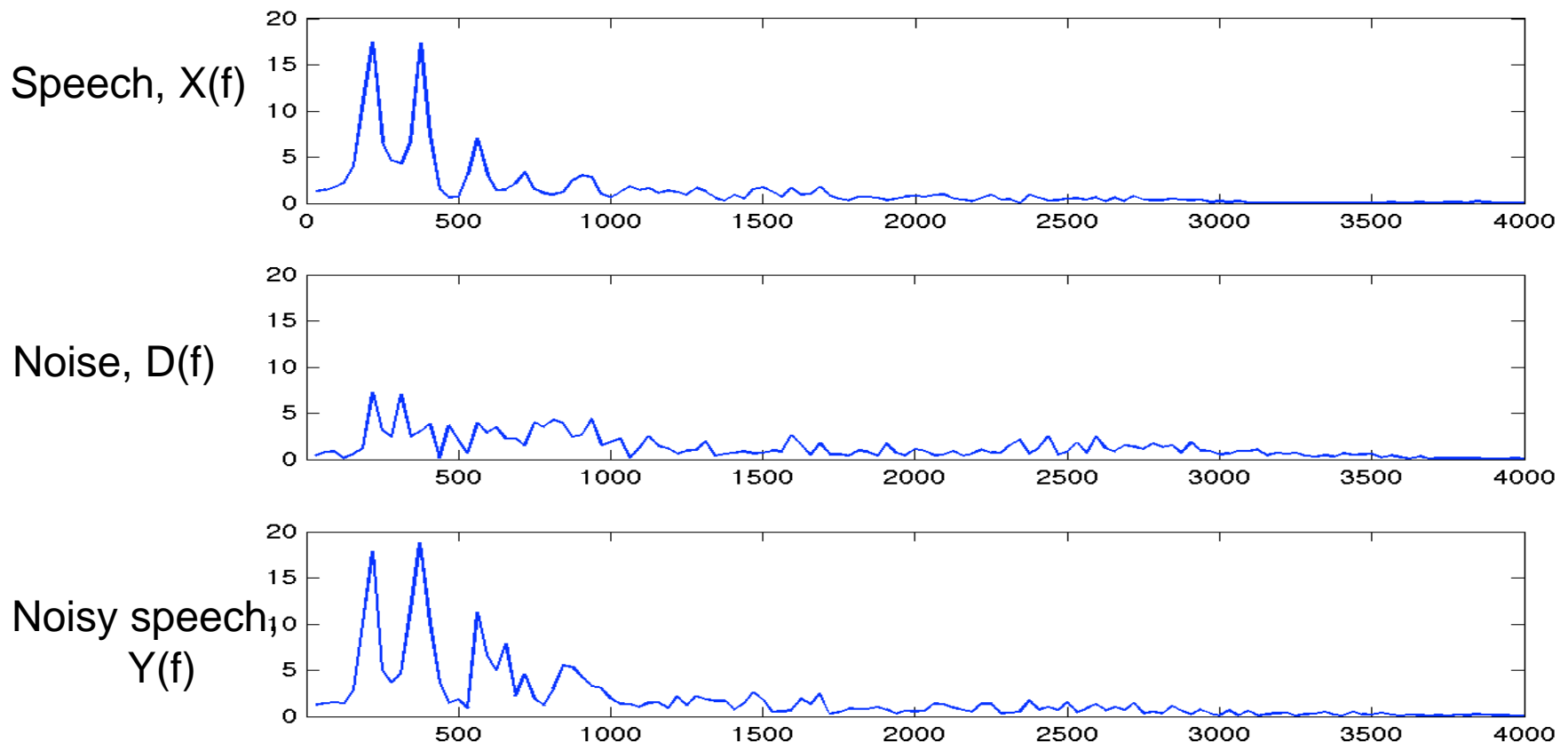


# Noisy Speech – Frequency-Domain

- Taking a discrete Fourier transform (DFT) to obtain the complex spectrum is a linear operation so the process is still additive

$$Y(f) = X(f) + D(f)$$

- Where  $Y(f)$ ,  $X(f)$  and  $D(f)$  are the complex spectra of the noisy speech, speech and noise



# Magnitude Spectrum

We don't usually deal with the complex spectrum, instead convert to magnitude spectrum and phase spectrum

$$Y(f) = X(f) + D(f)$$

Take magnitude

$$|Y(f)| = |X(f) + D(f)|$$

which can be assumed equal to (ignoring cross terms)

$$|Y(f)| = |X(f)| + |D(f)|$$

So in magnitude spectral domain, assume noisy speech magnitude spectrum is equal to speech magnitude spectrum plus noise magnitude spectrum

# Signal-to-Noise Ratio (SNR)

Signal-to-noise ratio (SNR) is measured in dBs (decibels), and is defined as

$$SNR(dB) = 10 \log_{10} \frac{P_{speech}}{P_{noise}}$$

Where speech power and noise power are measured as

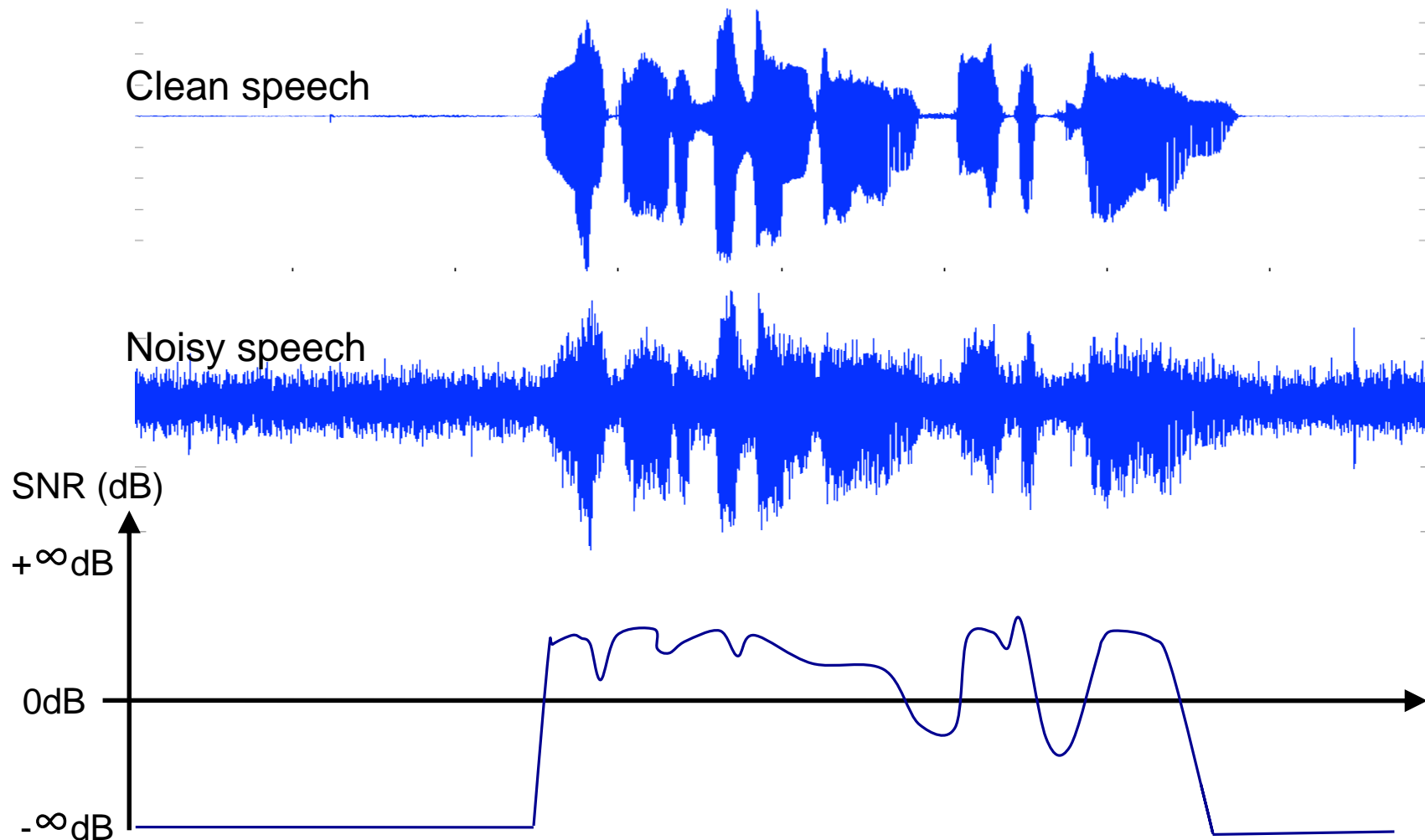
$$P_{speech} = \frac{1}{N} \sum_{n=0}^{N-1} x(n)^2 \quad P_{noise} = \frac{1}{N} \sum_{n=0}^{N-1} d(n)^2$$

So the SNR can be rewritten as

$$SNR(dB) = 10 \log_{10} \frac{\frac{1}{N} \sum_{n=0}^{N-1} x(n)^2}{\frac{1}{N} \sum_{n=0}^{N-1} d(n)^2} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} d(n)^2}$$

# SNR Across a Speech Utterance

Noisy speech contaminated by noise of approximately constant power



SNR varying locally across the utterance



# Filtering-Based Enhancement –Time-Domain

In the time-domain, clean speech signal can be recovered

$$x(n) = y(n) - d(n)$$

**Problem** – Normally we don't know the noise signal,  $d(n)$

Instead, we need to make an estimate of the noise and subtract this from the noisy speech to give enhanced speech

$$\hat{x}(n) = y(n) - \hat{d}(n)$$

Obtaining such a noise estimate in the time domain is not easy  
– better to transfer to the frequency-domain for noise estimation

Some noise reduction systems are operated in the time-domain but typically they can perform with a second microphone

# Filtering-Based Enhancement –Frequency Domain

A method to obtain enhanced speech by subtracting an estimate of the noise magnitude from the noisy speech magnitude in the frequency domain is known as **spectral subtraction (SS)**

$$\left| \hat{X}(f) \right| = \left| Y(f) \right| - \left| \hat{D}(f) \right|$$

We still need a noise estimate – some practical methods exist for estimation

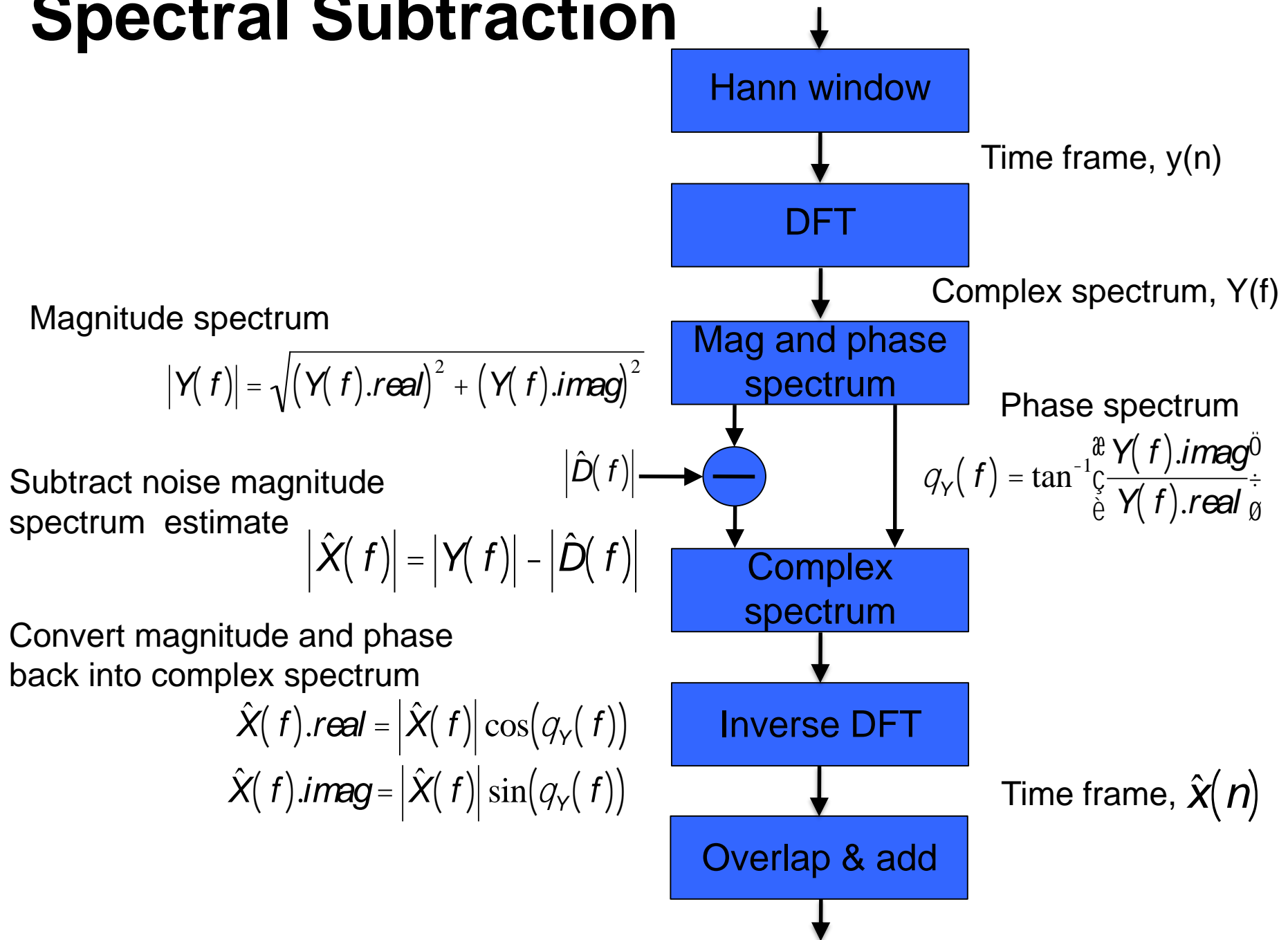
Speech enhancement typically outputs a time-domain speech signal that can be played out

Need to transform the enhanced magnitude spectrum back into a time-domain signal using the inverse DFT

Inverse DFT is applied to the complex spectrum, so need to go from magnitude spectrum and phase spectrum to complex spectrum before IDFT can be applied

Most methods combine enhanced magnitude spectrum with the noisy phase – ear not sensitive to phase

# Spectral Subtraction



# Problem with Spectral Subtraction

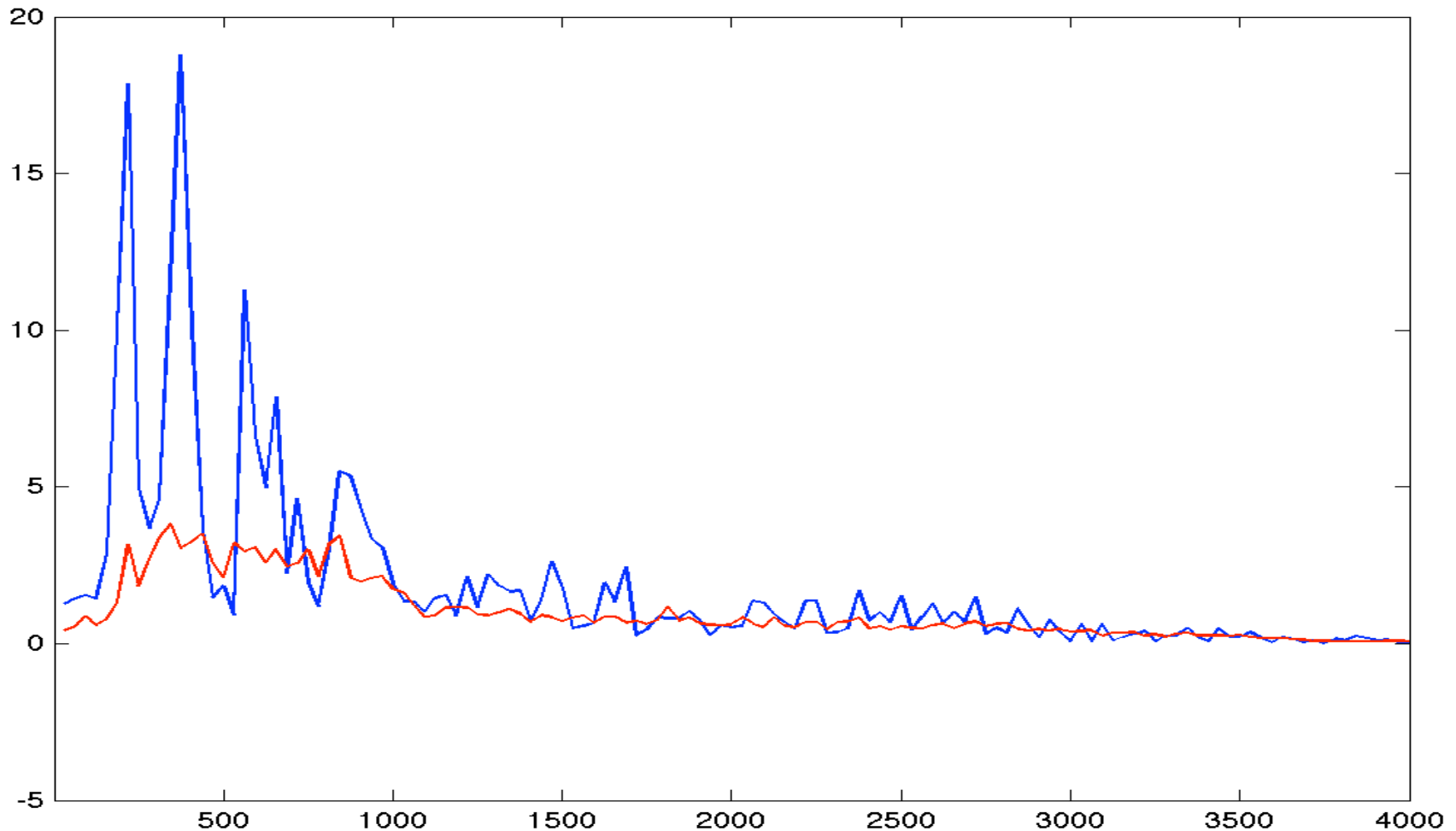
- Considering the spectral subtraction equation

$$|\hat{X}(f)| = |Y(f)| - |\hat{D}(f)|$$

- What happens if  $|\hat{D}(f)| > |Y(f)|$  i.e. noise magnitude estimate is greater than the noisy speech magnitude
- Cannot have a negative magnitude spectrum!
- Need to employ post-processing methods to ensure that this does not happen
  1. Half-wave rectification
  2. Spectral flooring
  3. Maximum attenuation threshold

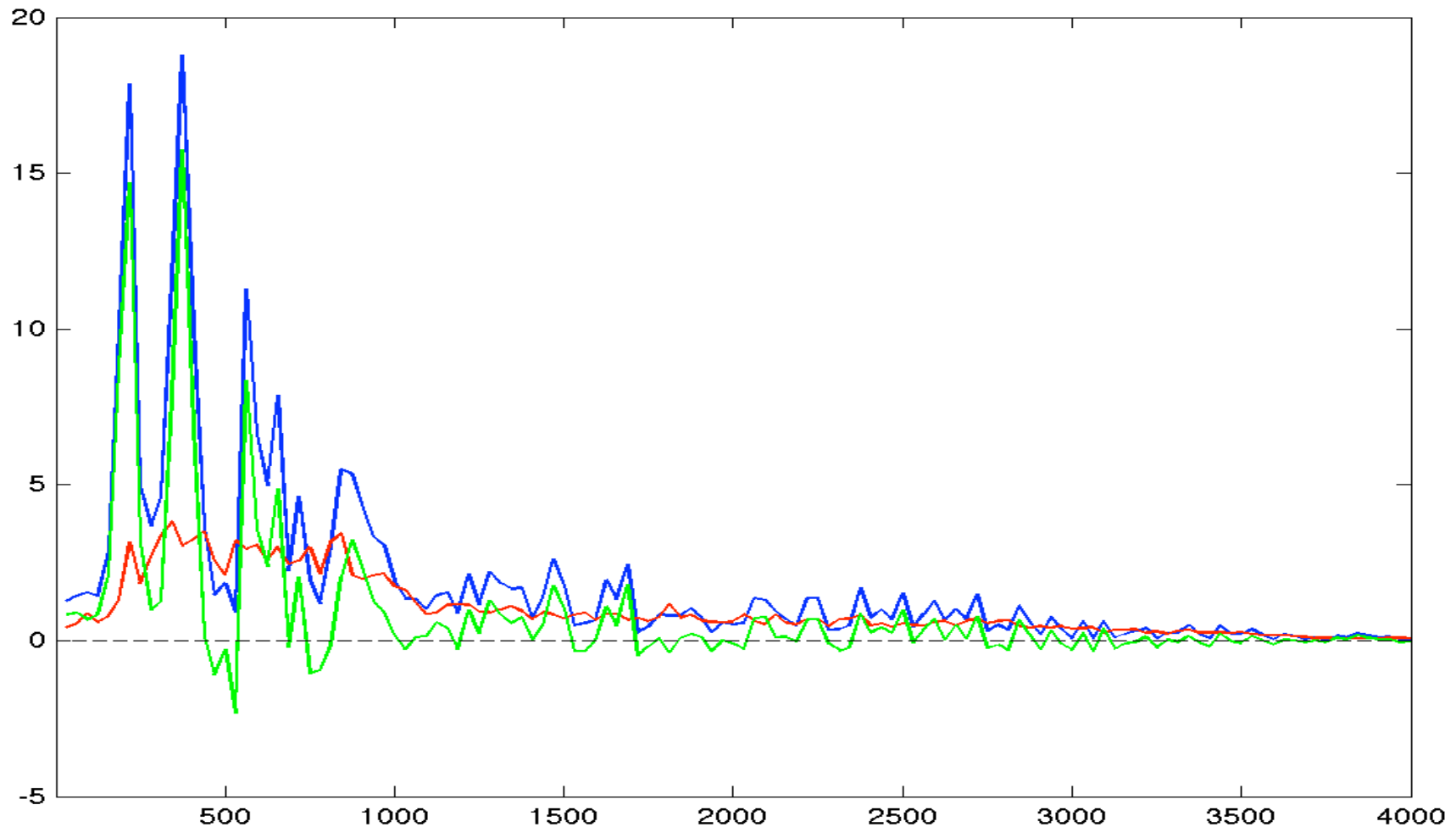
# Problem with Spectral Subtraction

- Consider the following example
  - Blue = Noisy speech
  - Red = Noise estimate
  - Green = Clean speech estimate



# Problem with Spectral Subtraction

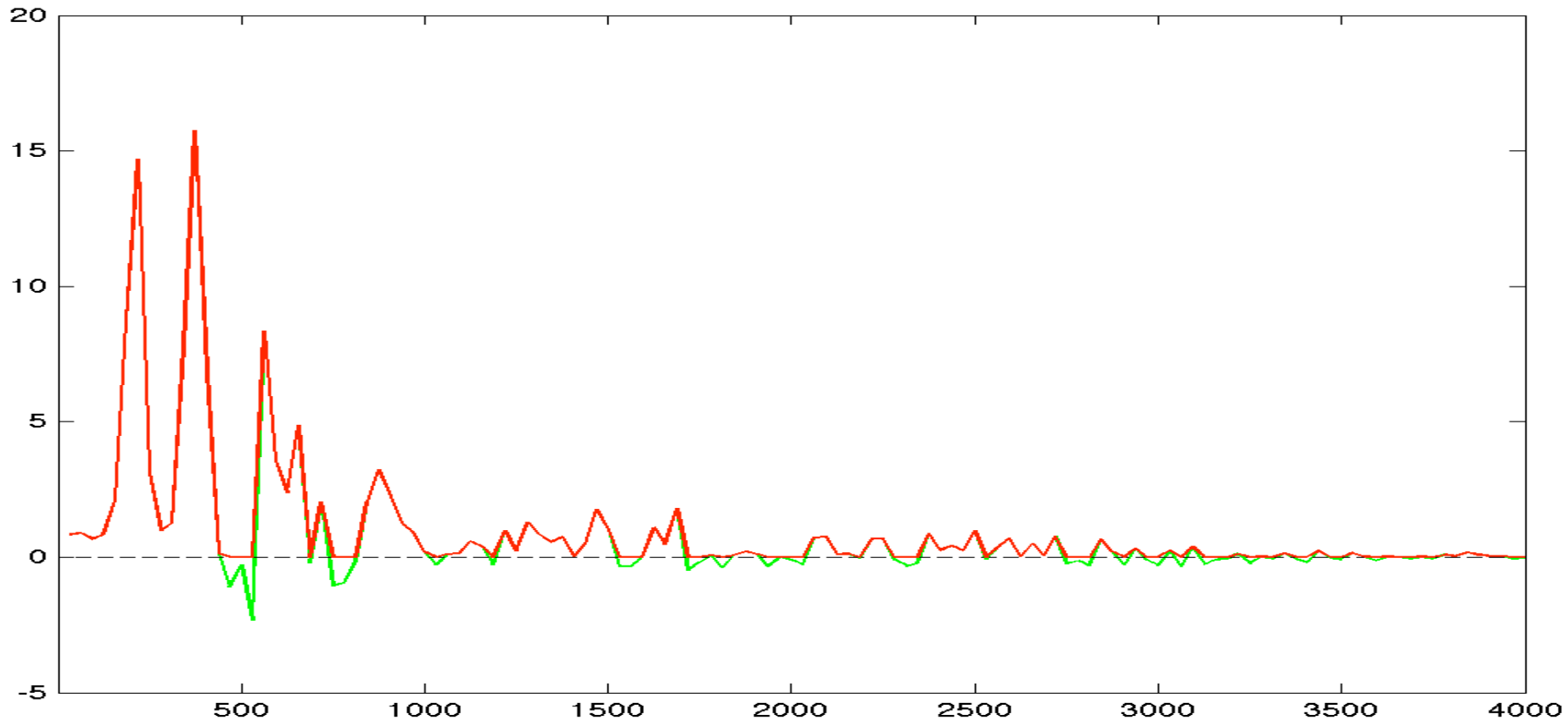
- Consider the following example
  - Blue = Noisy speech
  - Red = Noise estimate
  - Green = Clean speech estimate



# SS – Half-Wave Rectification

- Half-Wave Rectification

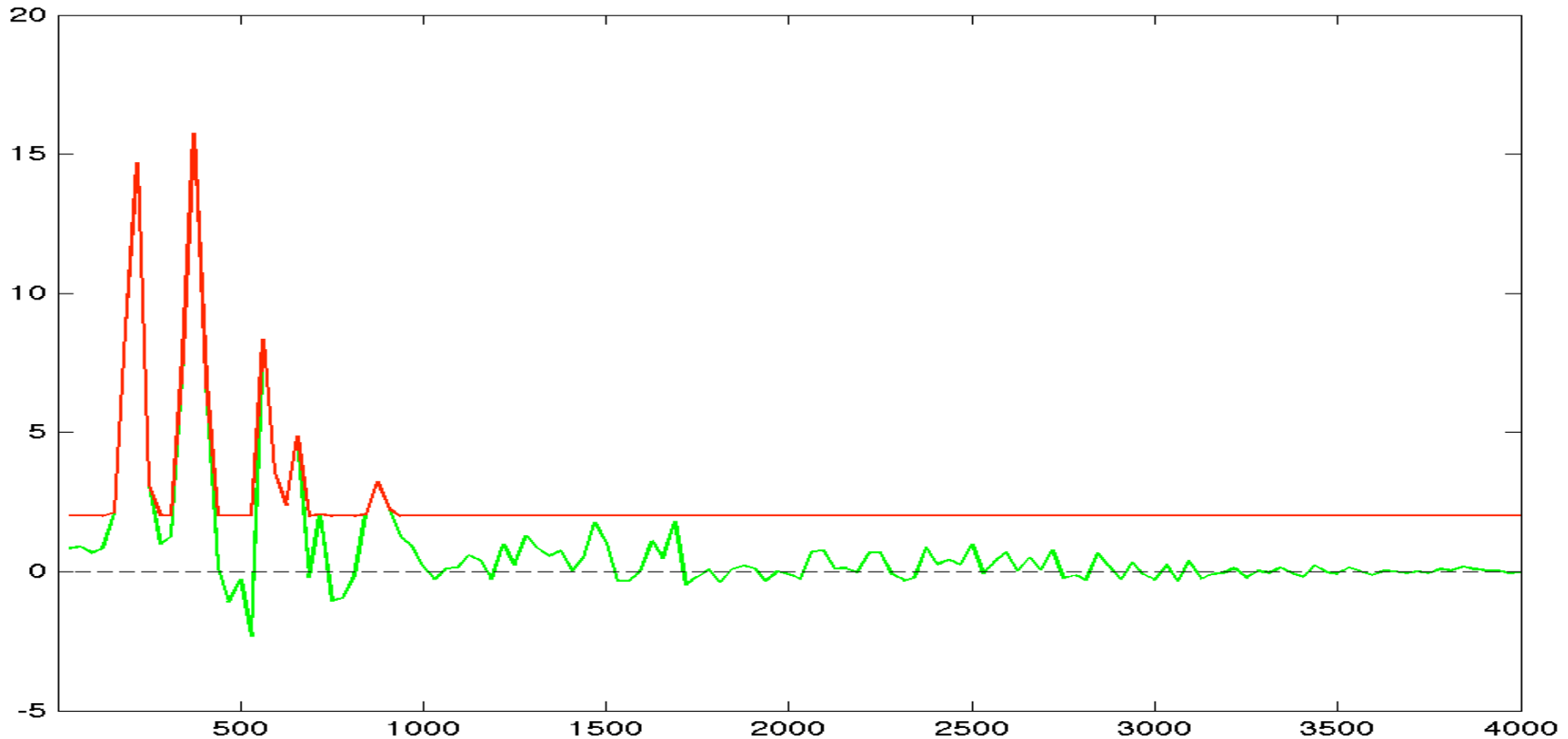
$$|\hat{X}(f)| = \begin{cases} |Y(f)| - |\hat{D}(f)| & \text{if } |Y(f)| \geq |\hat{D}(f)| \\ 0 & \text{otherwise} \end{cases}$$



# SS – Spectral Floor

- Spectral flooring to a noise floor (NF)

$$|\hat{X}(f)| = \begin{cases} |Y(f)| - |\hat{D}(f)| & \text{if } |Y(f)| - |\hat{D}(f)| > NF \\ NF & \text{otherwise} \end{cases}$$

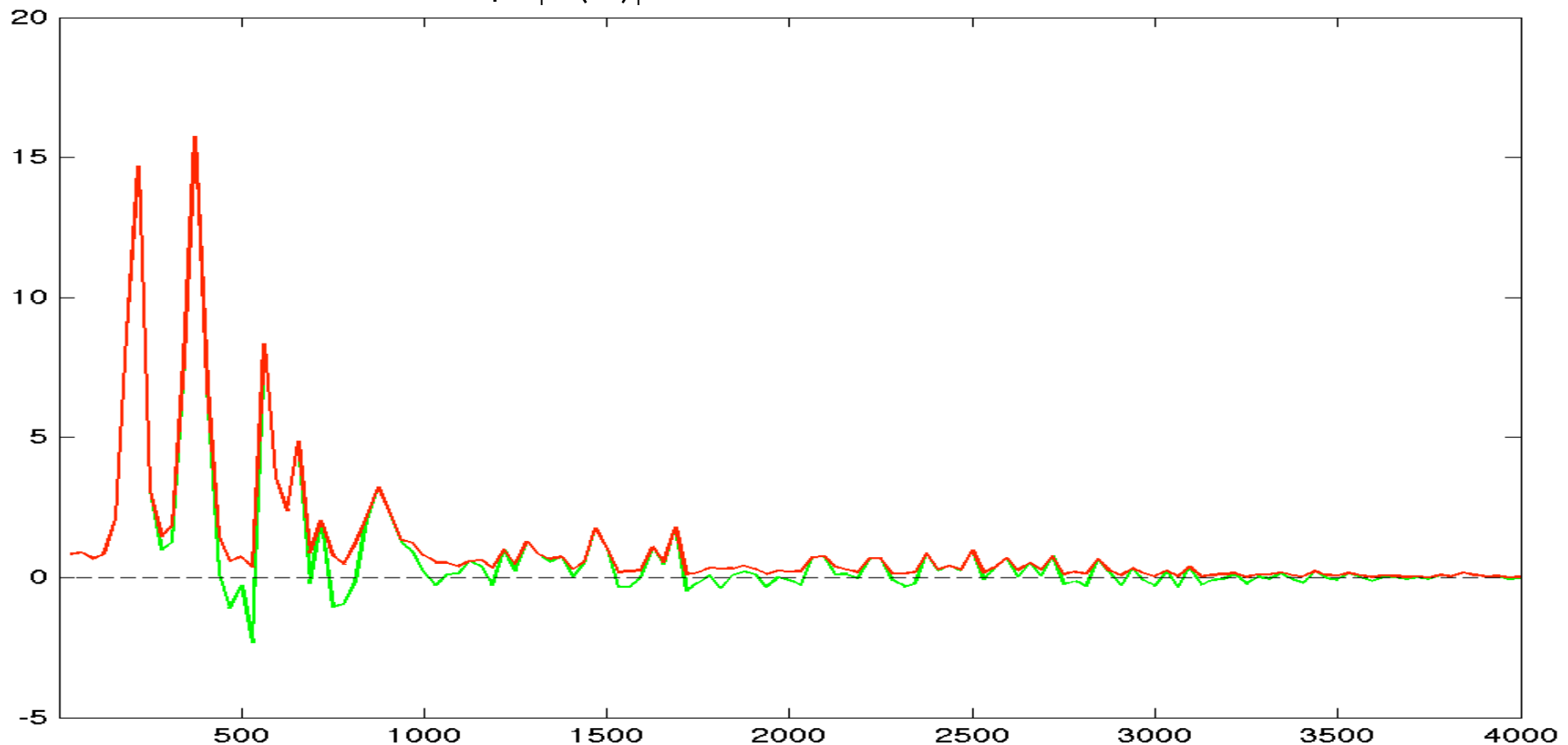




# SS – Maximum Attenuation Threshold

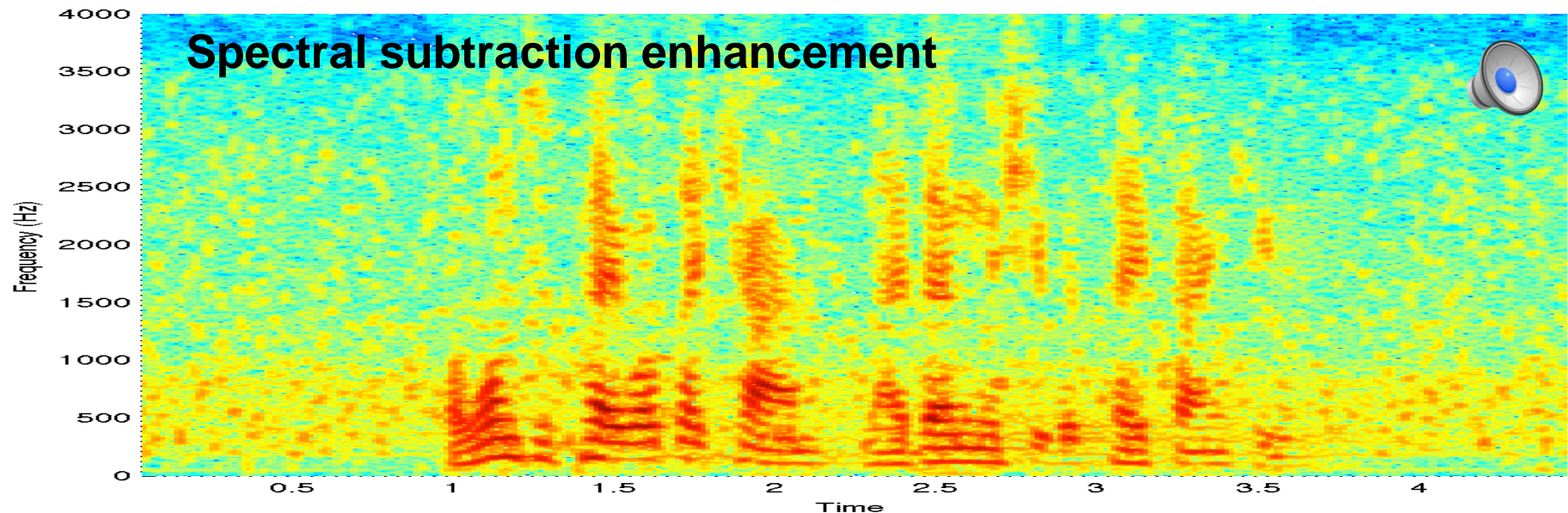
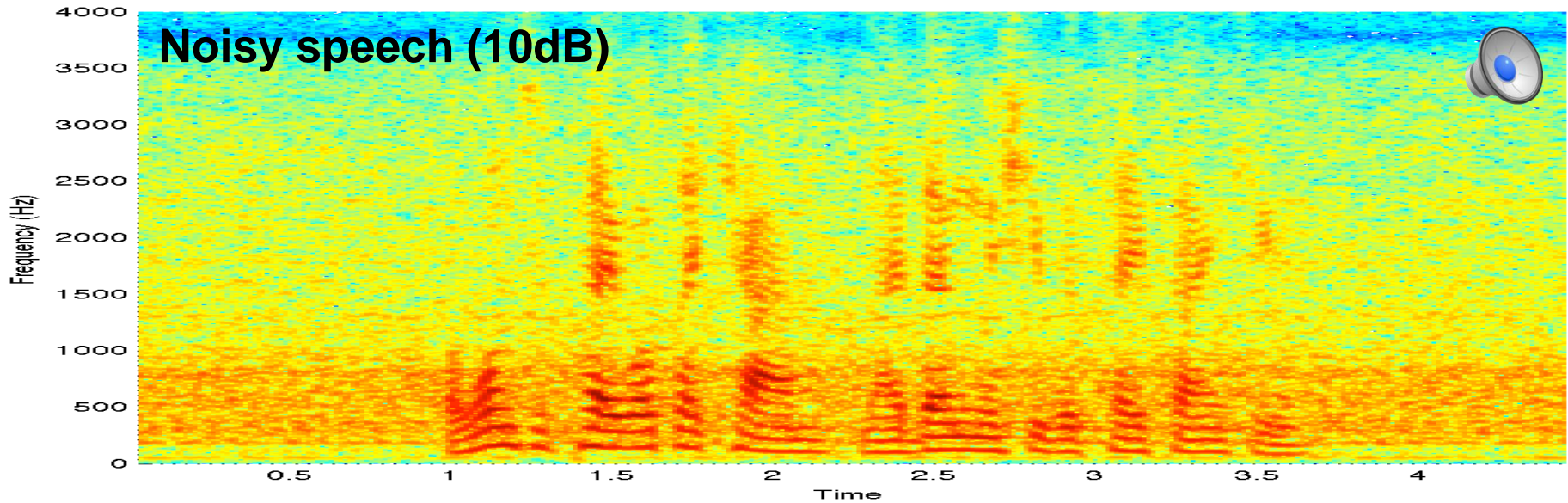
- Limit maximum attenuation to a gain of  $\beta$

$$|\hat{X}(f)| = \begin{cases} |Y(f)| - |\hat{D}(f)| & \text{if } |Y(f)| - |\hat{D}(f)| > b|Y(f)| \\ b|Y(f)| & \text{otherwise} \end{cases}$$



# SS – Residual & Musical Noise

- Problem with SS is residual noise and introduction of musical noise

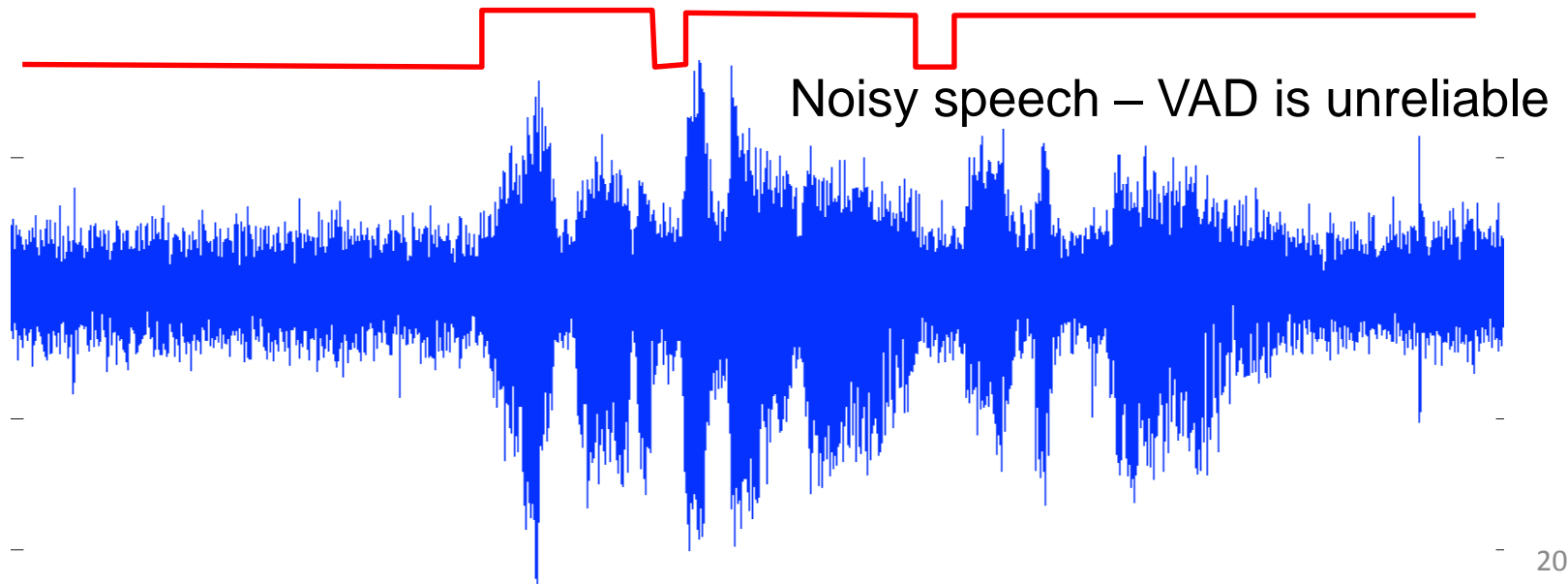


# Noise Estimation - VAD

- Spectral subtraction (and most speech enhancement algorithms) require an estimate of the noise signal
- Many methods exist to obtain this – will consider voice activity detection (VAD)
- Idea is to identify periods of speech and non-speech and update noise estimate during periods of non-speech and use it during speech
- Most VADs use a measure of the signal energy to determine whether the signal is speech or non-speech
- Compute energy of short duration frames  $E_i = \frac{1}{N} \sum_{n=0}^{N-1} y_i(n)^2$
- Make VAD decision 
$$VAD_i = \begin{cases} 1 & \text{if } E_i > \gamma \\ 0 & \text{otherwise} \end{cases}$$
- Threshold  $\gamma$  determined from noise and speech energy levels

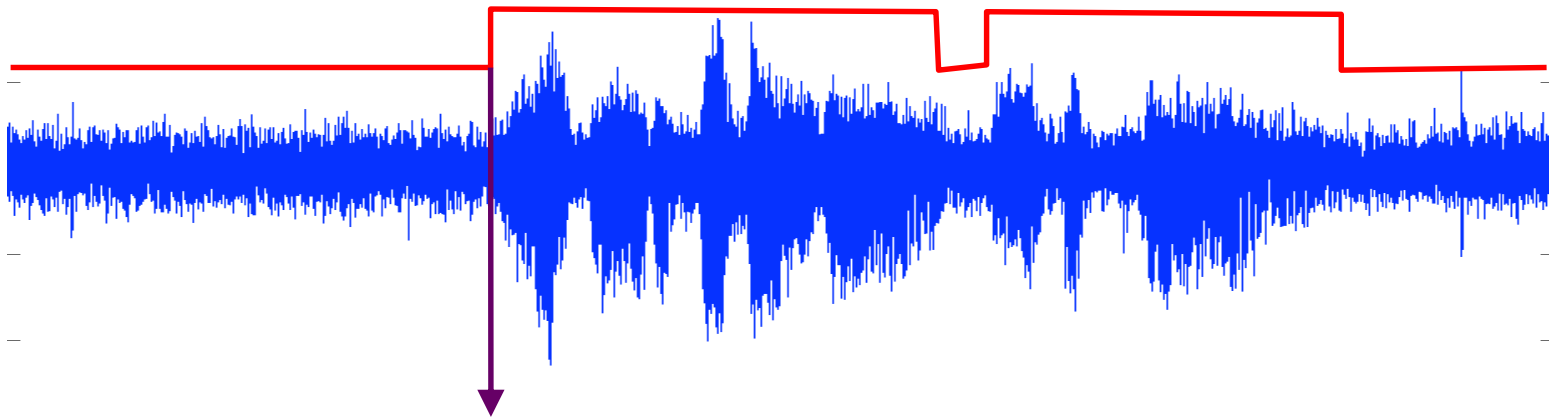
# Noise estimation - VAD

- Compute energy of short duration frames  $E_i = \frac{1}{N} \sum_{n=0}^{N-1} y_i(n)^2$
- Make VAD decision 
$$VAD_i = \begin{cases} 1 & \text{if } E_i > \gamma \\ 0 & \text{otherwise} \end{cases}$$
- Threshold  $\gamma$  determined from noise and speech energy levels



# Noise Estimation

- Using the output of the VAD, audio frames that are labelled as non-speech can be averaged and an estimate of the noise made



$$|\hat{D}(f)| = \frac{1}{N} \mathring{a} \sum_{i=0}^{N-1} |Y_i(f)|$$

- Noise estimate updating stops during speech periods (where enhancement takes place) and resumes again in non-speech periods

# Other Methods of Filtering-Based Enhancement

## Wiener Filtering

Optimised linear filter between the complex spectra of the noisy and clean speech in terms of mean-square error.

$$\hat{X}(f) = H(f)Y(f)$$

$$H(f) = \frac{\xi(f)}{\xi(f) + 1}$$

$$\xi(f) = \frac{E[|X(f)|^2]}{E[|D(f)|^2]} \quad \text{a priori SNR}$$

$$\xi(f) = \alpha \frac{|\hat{X}^{(-)}(f)|^2}{|\hat{D}^{(-)}(f)|^2} + (1 - \alpha) \left( \frac{|Y(f)|^2}{|\hat{D}(f)|^2} + 1 \right) \quad \text{a posteriori SNR}$$

$|\hat{X}^{(-)}(f)|^2$  : Power of the enhanced speech at the previous frame

$|\hat{D}^{(-)}(f)|^2$  : Power of the noise estimate at the previous frame

# Other Methods of Filtering-Based Enhancement

## ***Log MMSE Estimator***

Nonlinear filter between the magnitude of the clean and noisy speech

$$|\hat{X}(f)| = G(\xi(f), \gamma(f)) |Y(f)|$$

$$\xi(f) = \frac{E[|X(f)|^2]}{E[|D(f)|^2]} \quad \text{a priori SNR}$$

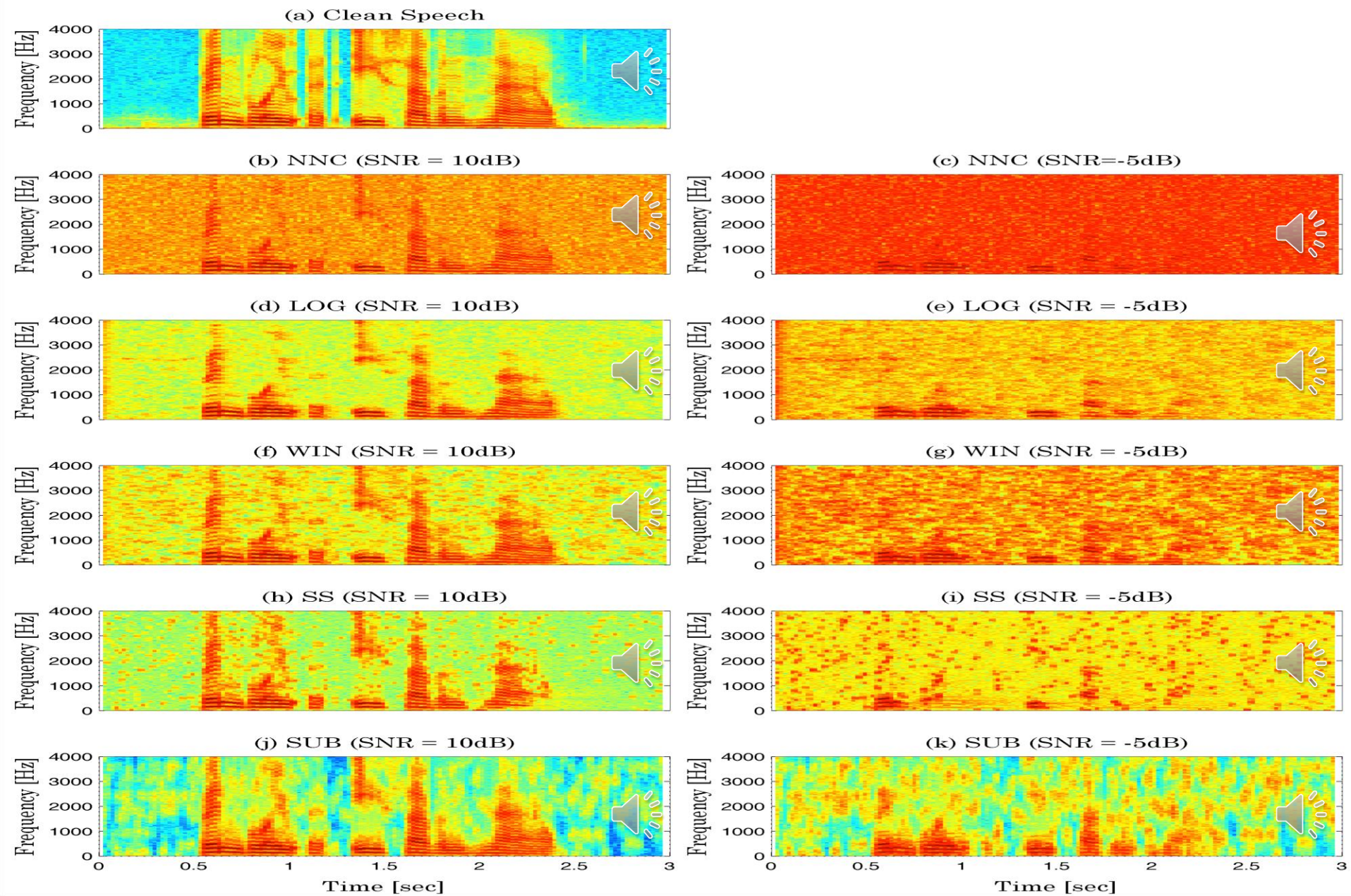
$$\gamma(f) = \frac{|Y(f)|^2}{E[|D(f)|^2]} \quad \text{a posteriori SNR}$$

## ***Subspace Method***

Noisy speech is transformed into a new space that comprises speech and noise subspaces by using Singular Value Decomposition (SVD) so that the signal in the speech subspace can be taken as the enhanced speech signal

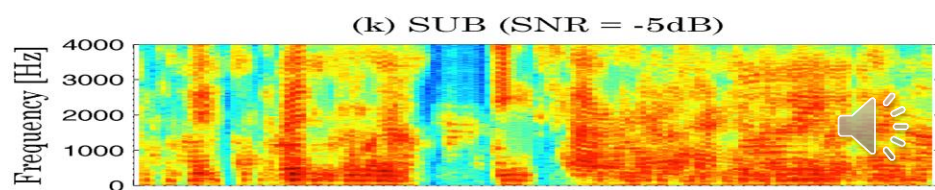
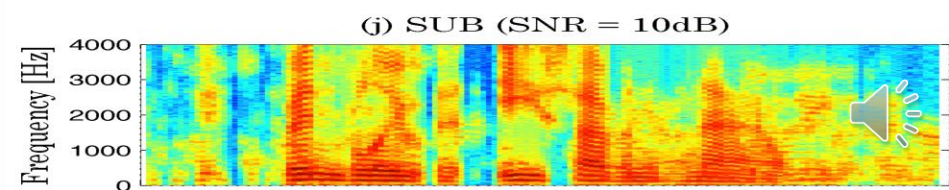
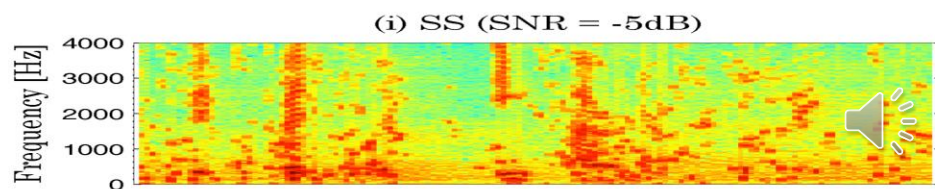
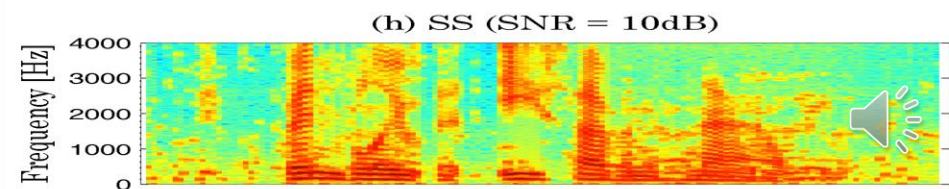
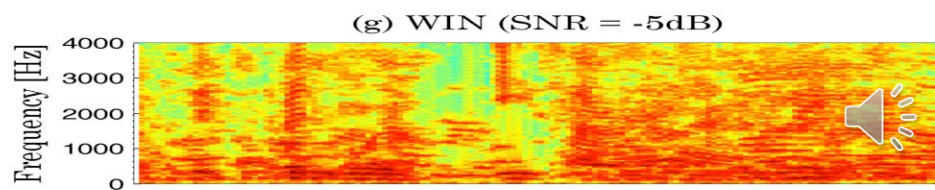
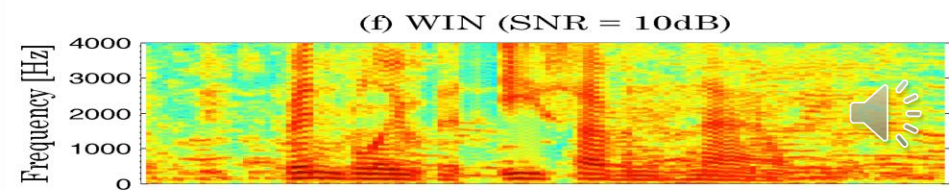
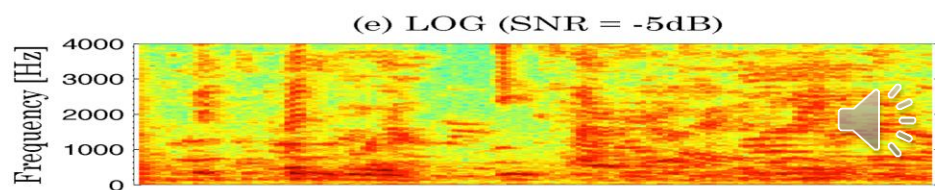
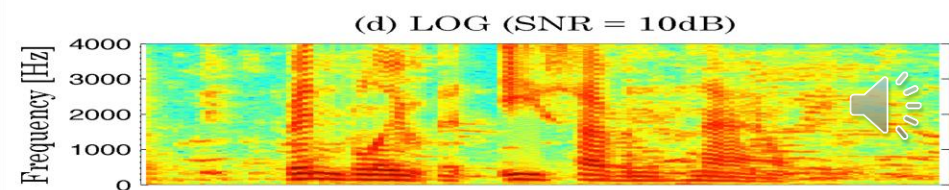
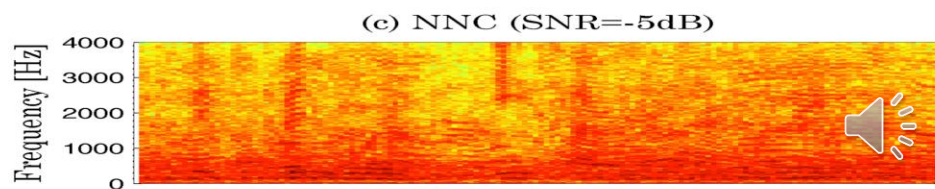
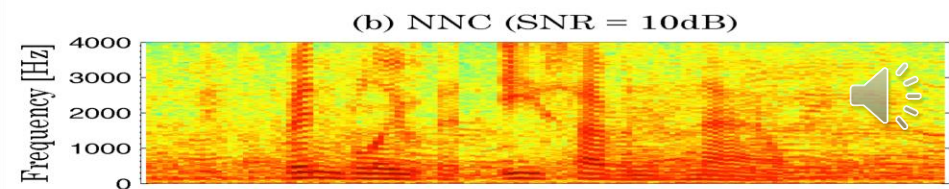
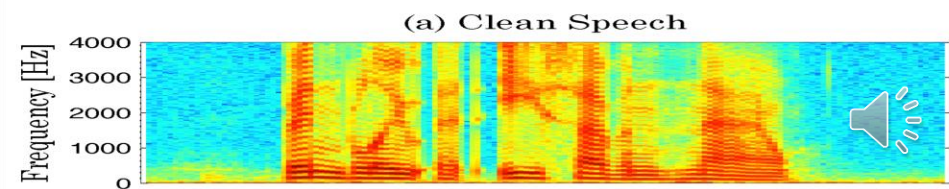


# Other Methods of Filtering-Based Enhancement





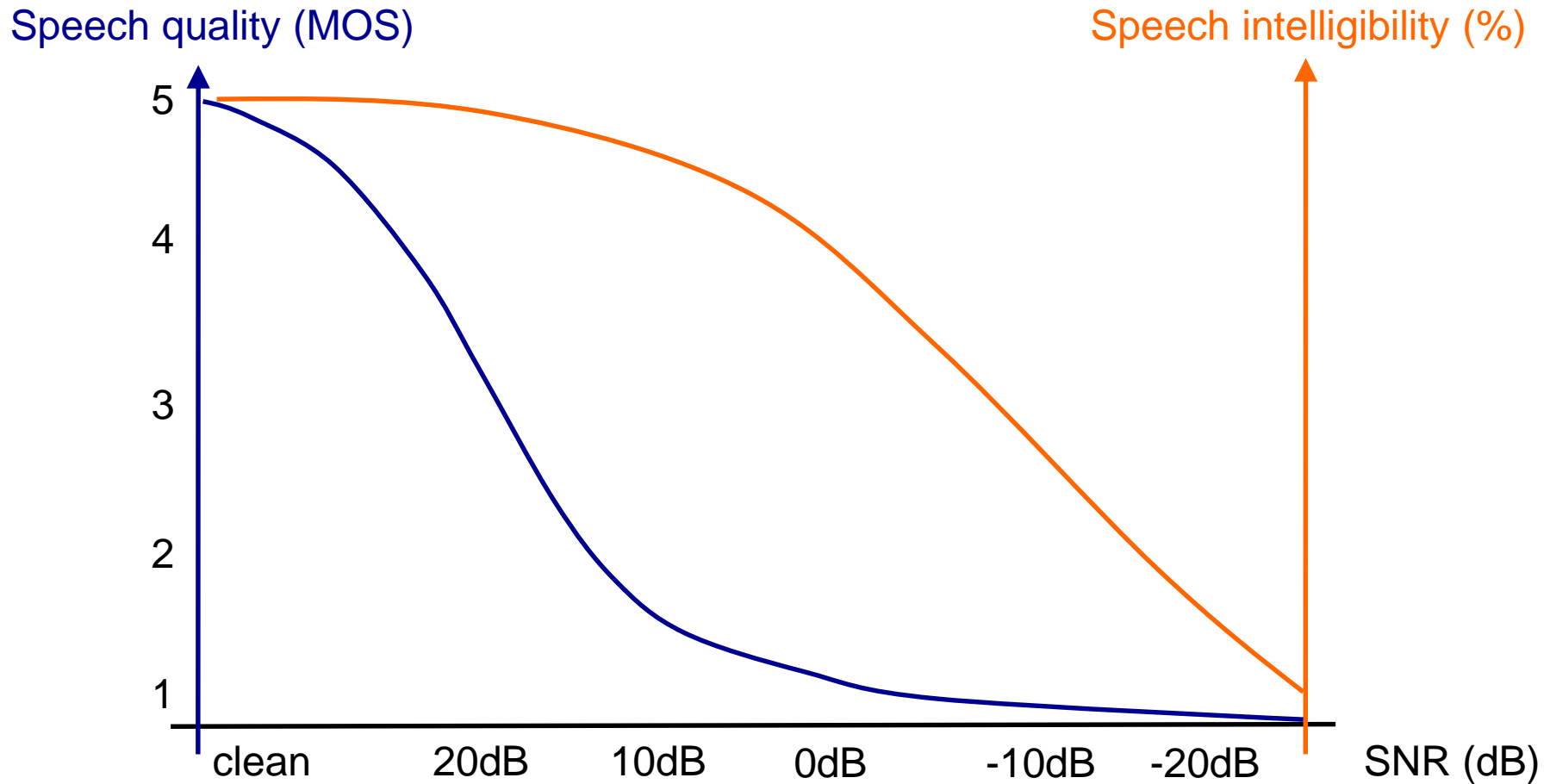
# Other Methods of Filtering-Based Enhancement



# Measuring Speech Quality / Intelligibility

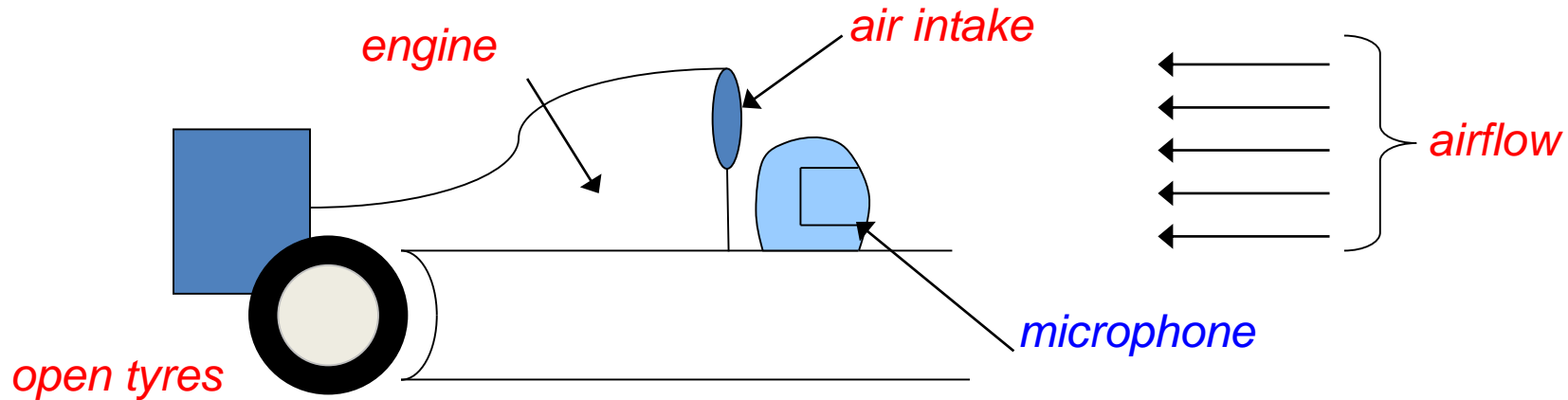
- Speech quality
  - Measure of how good the speech sounds
  - Measured subjectively using human listeners
  - Mean opinion score (MOS), scale of 1-5
  - Measured objectively using computer
  - PESQ, SNR, LLR
- Speech intelligibility
  - Measure of how understandable the speech is
  - Measured subjectively using human listeners – Recognition test
  - Measured objectively using computer
  - STOI, NCM, CSII

# Measuring speech quality/intelligibility



# Case Study – Speech Enhancement for Formula 1 Motor Racing

- Primary noise sources are **engine**, **airflow** and **tyres**



- Noisy speech signal,  $y(n)$ , can be modelled as,

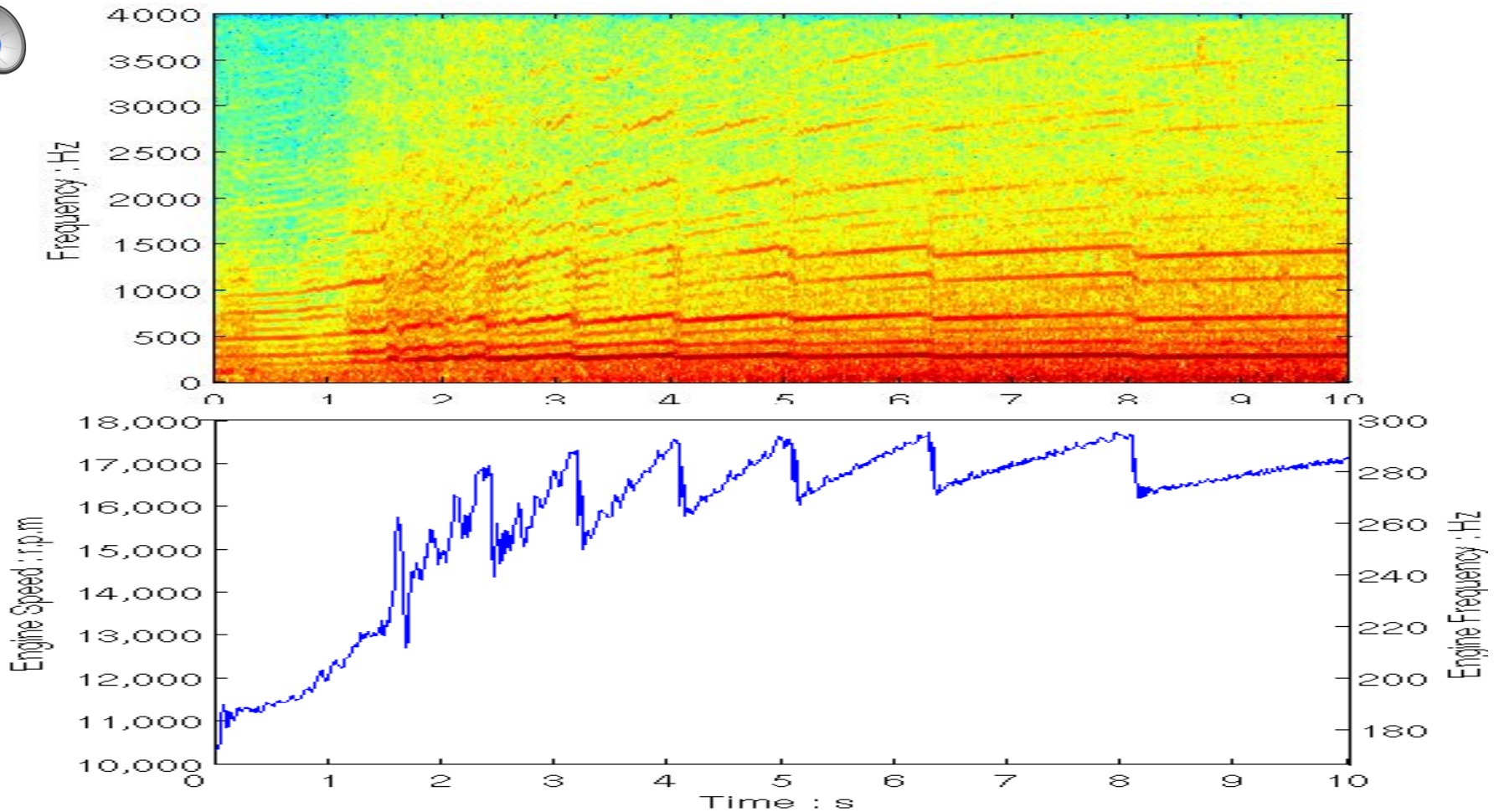
$$y(n) = x(n) + d_E(n) + d_{TA}(n)$$

- Sensor measurements on the vehicle provide a data stream of parameters, sampled at 100Hz, that give:

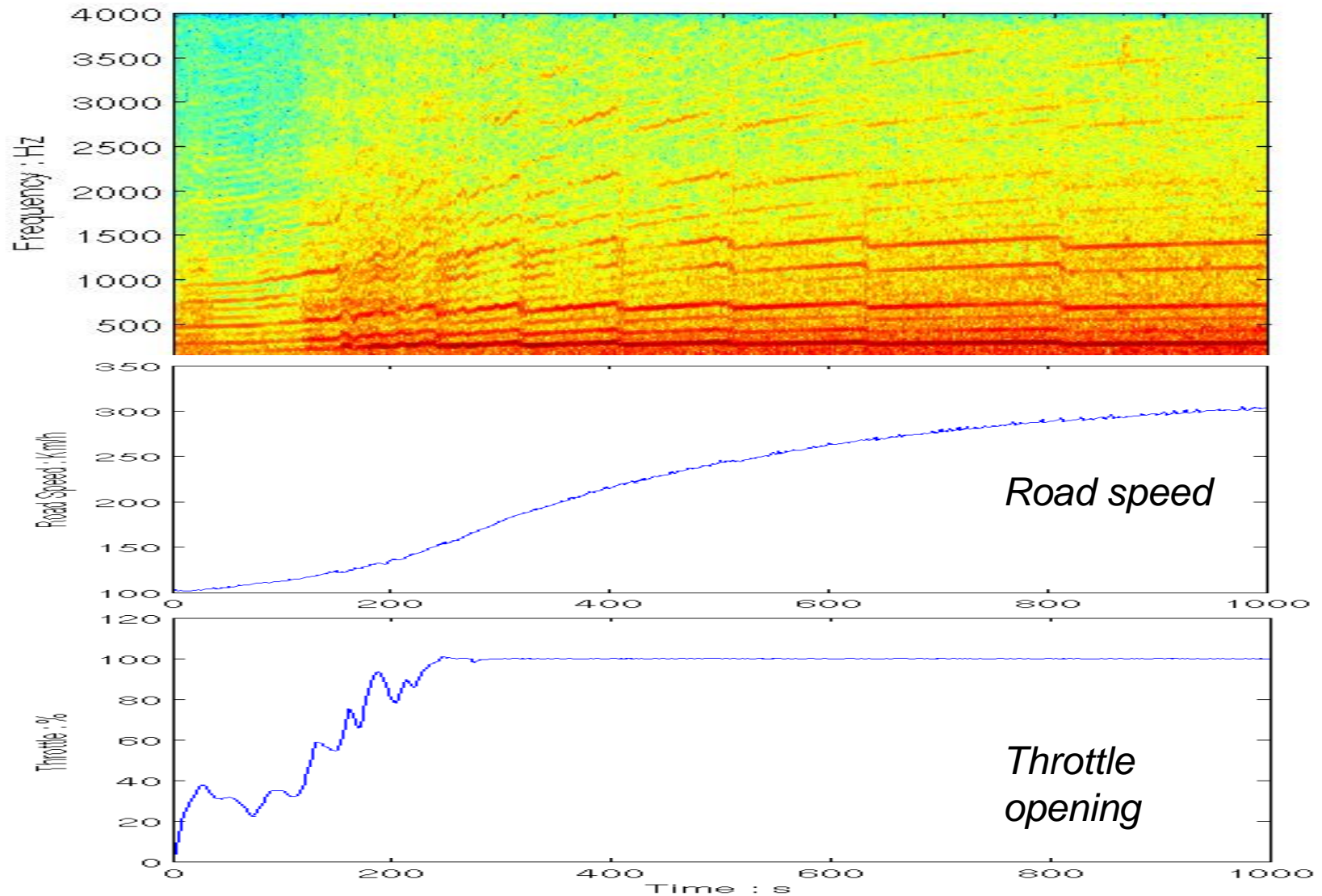
[*engine speed, road speed, throttle opening*]



# Case Study – Speech Enhancement for Formula 1 Motor Racing

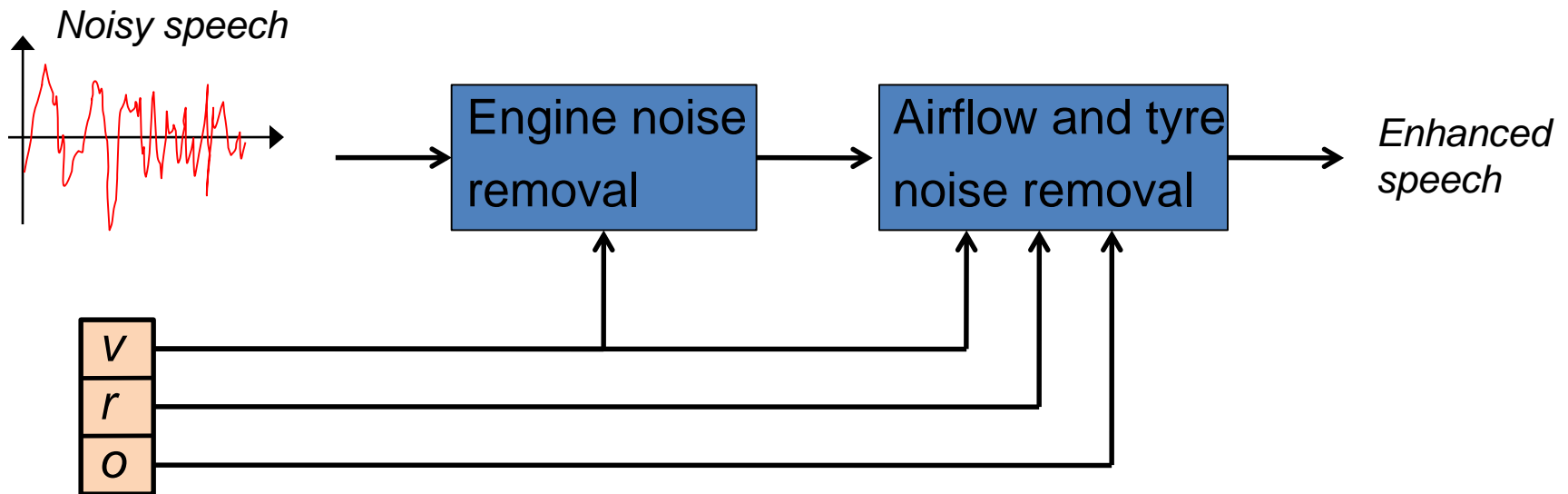


# Case Study – Speech Enhancement for Formula 1 Motor Racing



# Case Study – Speech Enhancement for Formula 1 Motor Racing

- Two-stage speech enhancement
  - Engine noise removal
  - Airflow and tyre noise removal



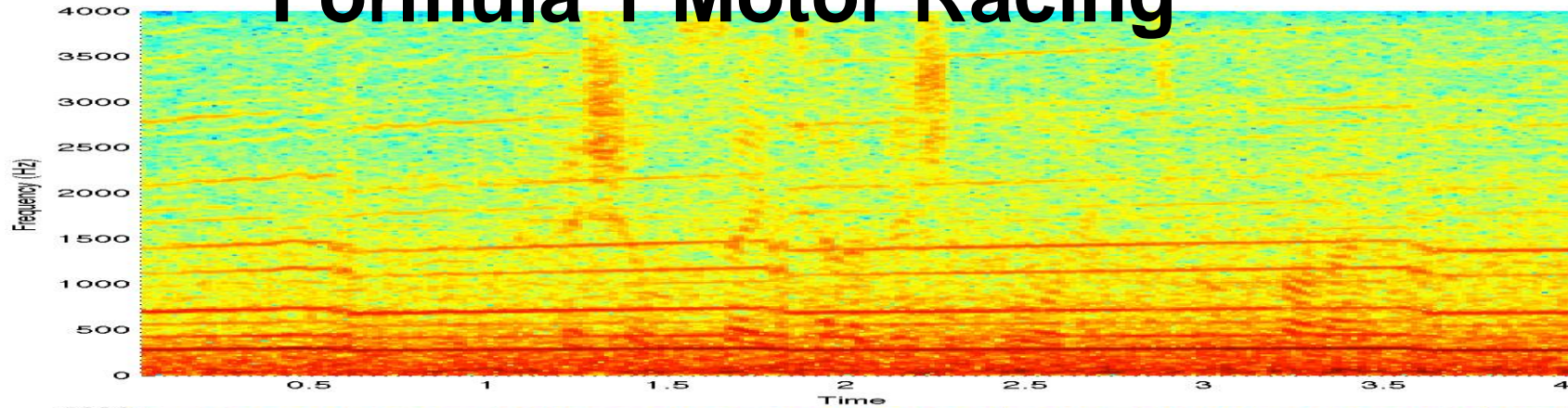
*datastream*

- *engine speed,  $v$*
- *road speed,  $r$*
- *throttle opening,  $o$*

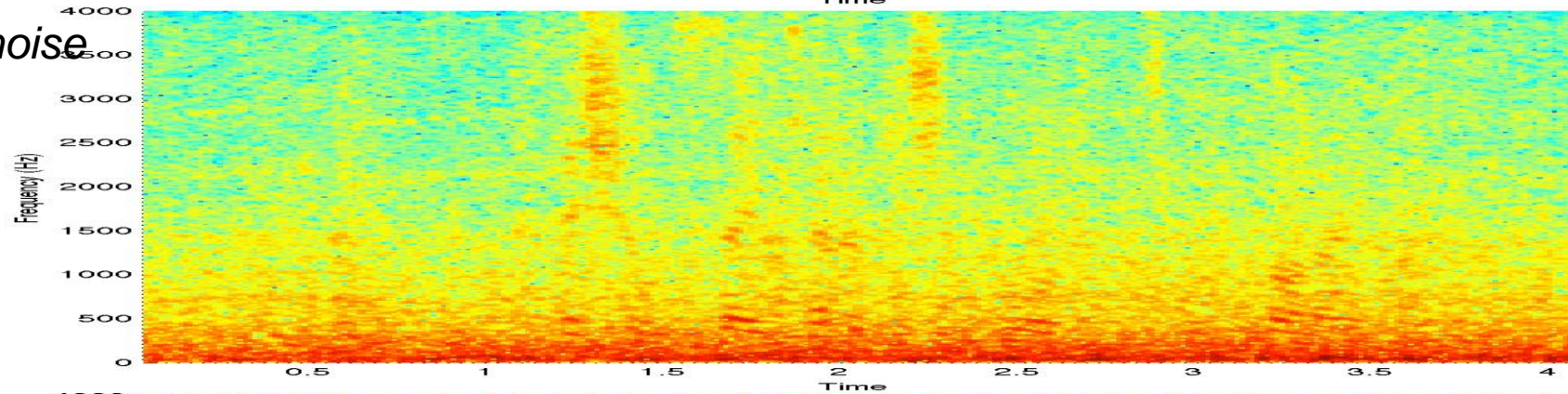


# Case Study – Speech Enhancement for Formula 1 Motor Racing

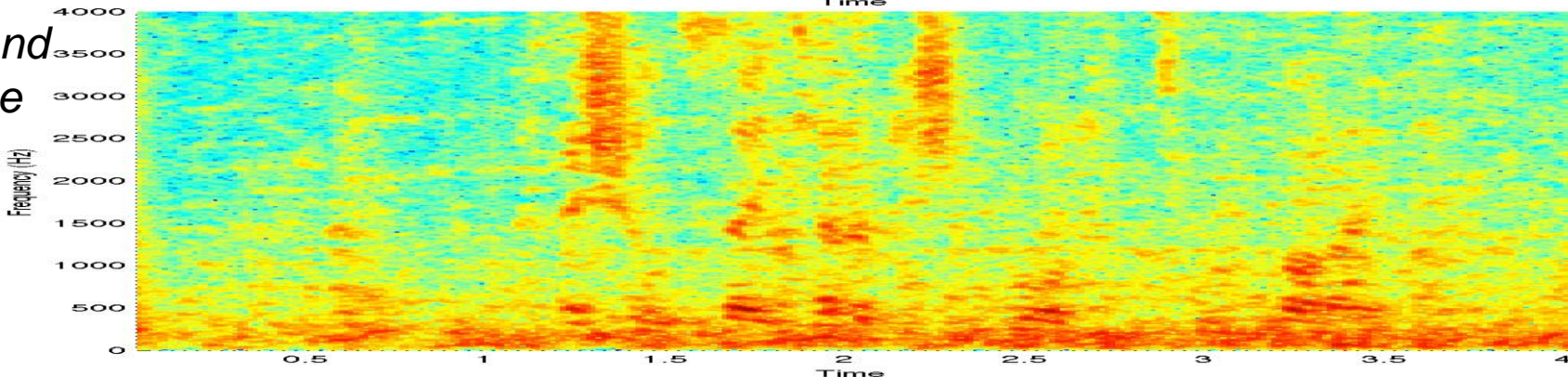
*Original*



*Engine noise  
removal*



*Airflow and  
tyre noise  
removal*

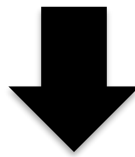




# Model-Based Speech Enhancement

Filtering-based speech enhancement cannot get rid of background noise especially at low SNRs

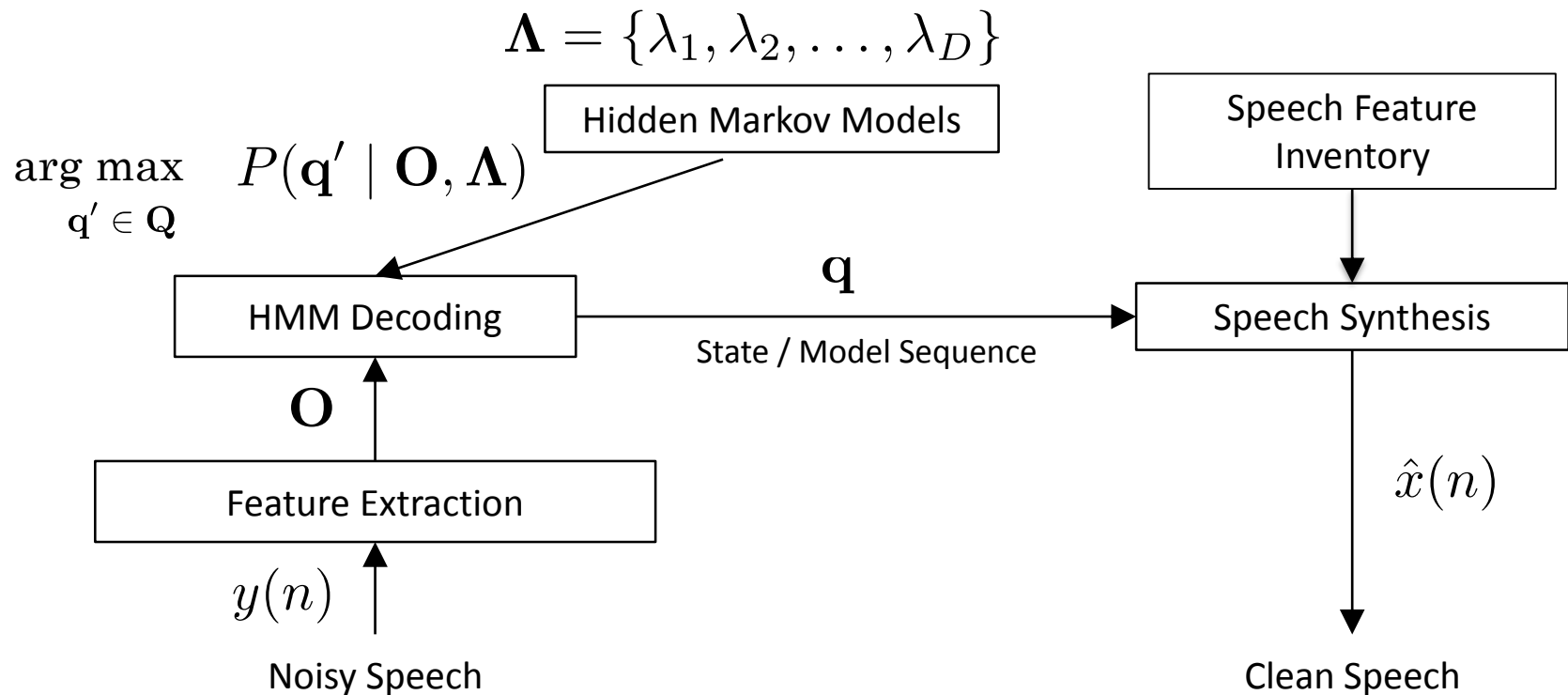
- Residual noise (Underestimation)
- Musical noise (Overestimation)



Consider model-based speech enhancement

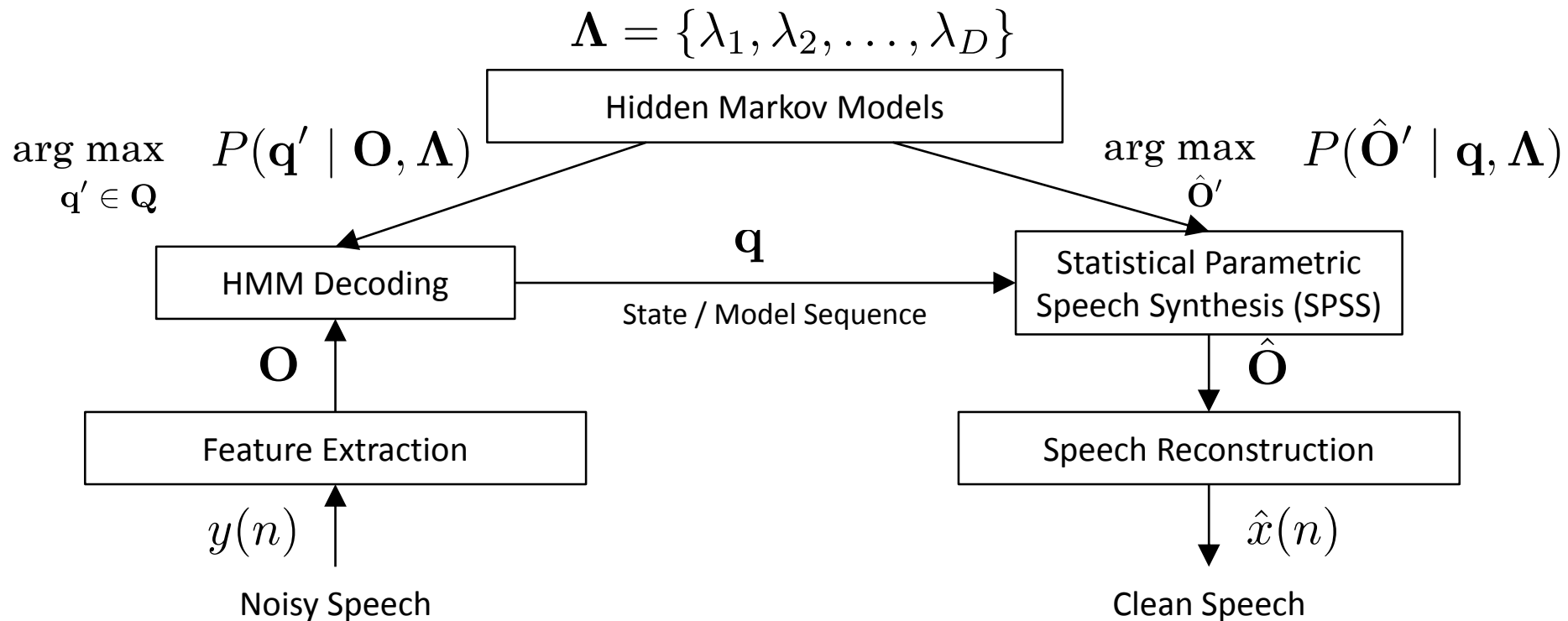
- Reconstruct background noise-free speech
- Statistical Models
- Speech synthesis (Speech production models)

# Model-Based Speech Enhancement



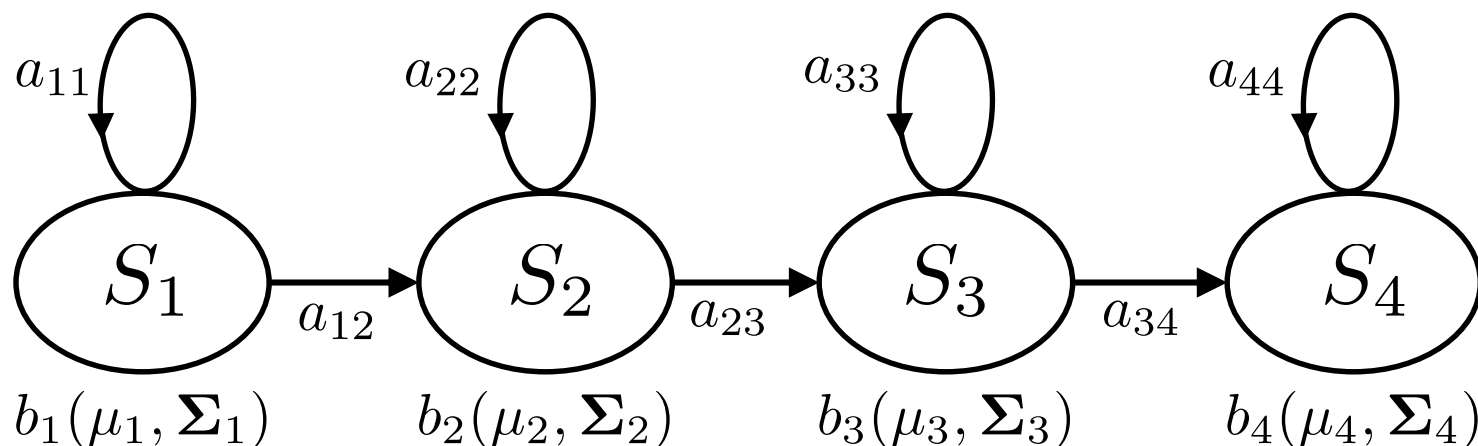
- Reconstruct clean speech
- Off-line process (Trained HMMs / Inventory)
- More computation
- Speaker Dependent

# Model-Based Speech Enhancement



- Get rid of a speech feature inventory
- More artificial speech
- More computation
- Speaker Independent (Model adaptation / i-vector)

# Hidden Markov Models



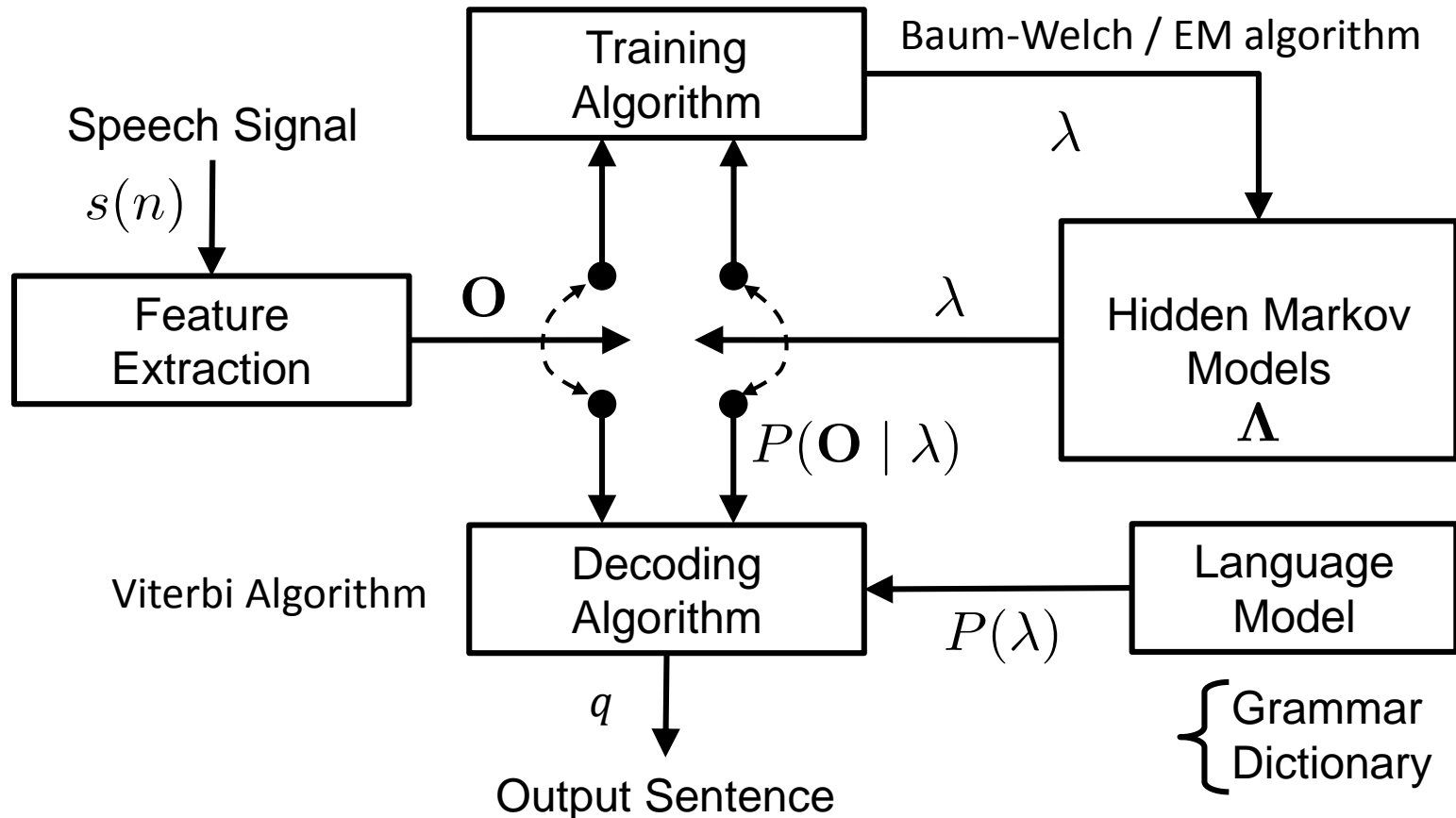
Left-to-Right Markov chain

$a_{ij}$  : Transition Probability from state  $i$  to state  $j$

$b_i$  : Distribution of outputs at state  $i$

- GMM-HMMs
- DNN-HMMs

# Training /Decoding Process



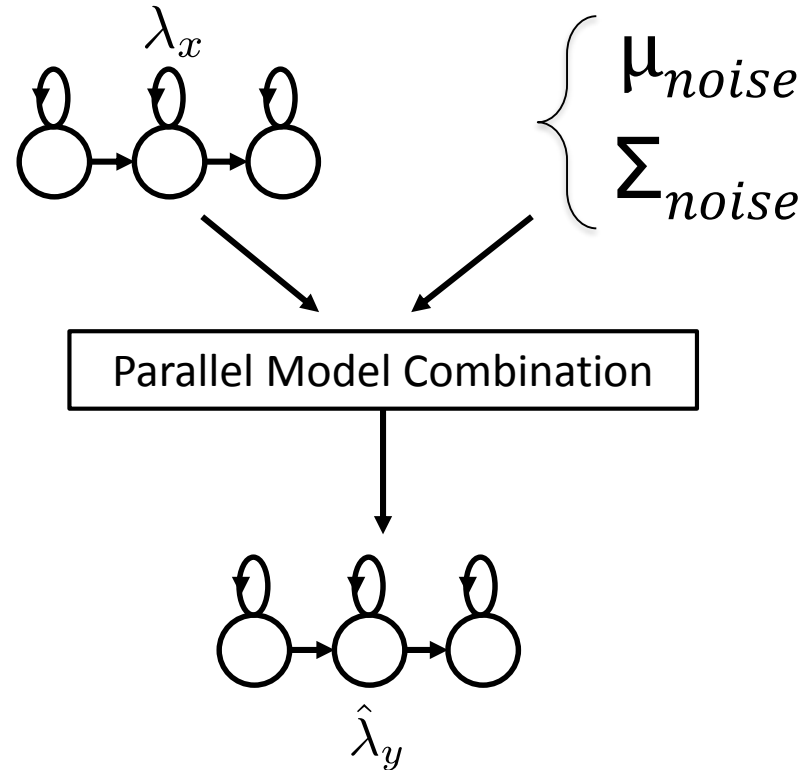
$$q = \operatorname{argmax}_{\lambda} \frac{P(\mathbf{O} | \lambda)P(\lambda)}{P(\mathbf{O})}$$

# Model Adaptation

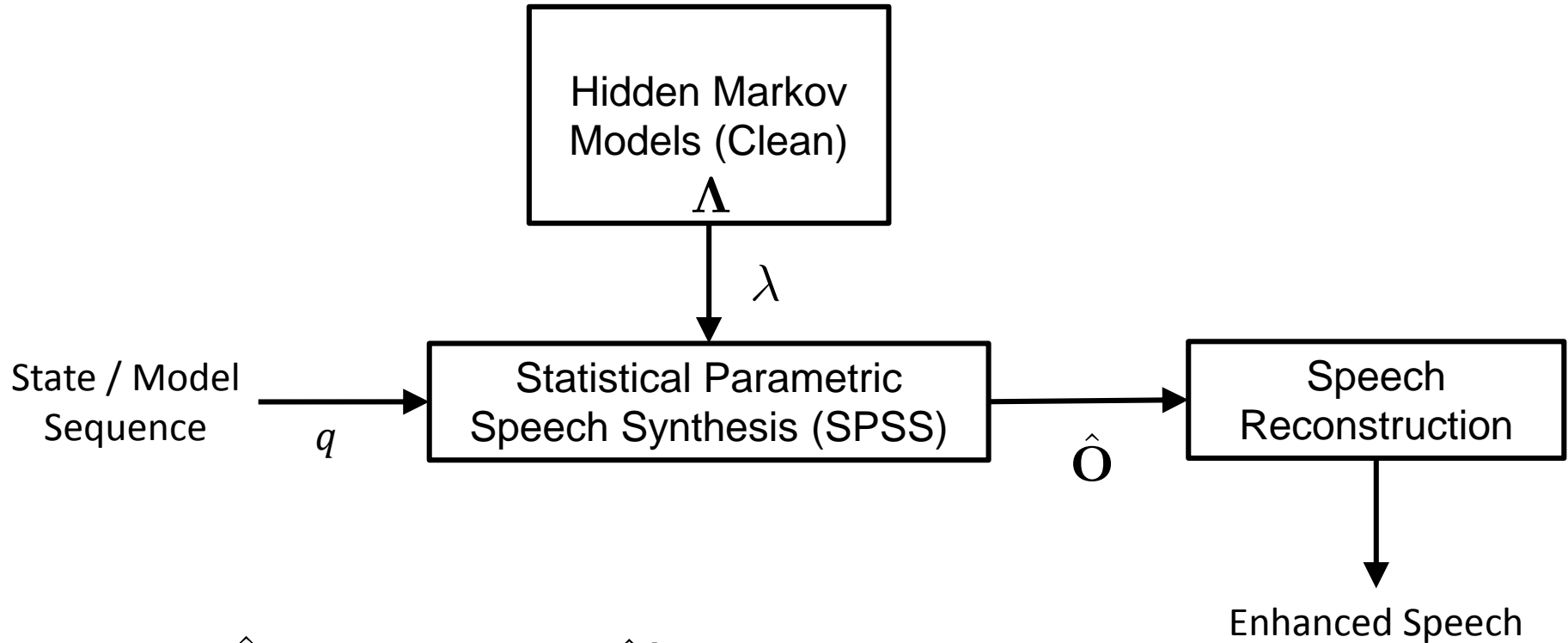
The lower SNRs, the lower decoding accuracy it gives.

Decoding errors makes enhanced speech different from the original speech.

- Model Adaptation
- DNN-HMMs











# Synthesis Process



$$\hat{\mathbf{O}} = \arg \max P(\hat{\mathbf{O}}' \mid q, \lambda)$$

Refer to  
H. Zen, K. Tokuda and A. W. Black, "Spectral Parametric Speech Synthesis", Speech Communication, 2009

# Enhanced Speech

	Male / White	Female / Babble
Noisy (-5dB)		
Filtering-Based (Log MMSE)		
Model-Based		
Clean (Ref)		



# Summary

- Noise process can be assumed to be additive in time and frequency domains
- Amount of noise can be measured by the SNR
- Spectral subtraction is a relatively simple method of speech enhancement and works by subtracting an estimate of the noise
- Problem arises in situations where noise estimate leads to negative magnitude
- Filtering-based enhancement introduces musical and residual noise onto the enhanced speech signal
- Simple noise estimation method is to use a VAD
- Model-based speech enhancement synthesises background noise-free speech as the enhanced speech – off-line process / artefacts / decoding errors