Study And Evaluation of Innovative Algorithms for Voice Quality Enhancement in Speech Signals Encoded Using ACELP (Algebraic Code Excited Linear Prediction)

Daniele GIACOBELLO

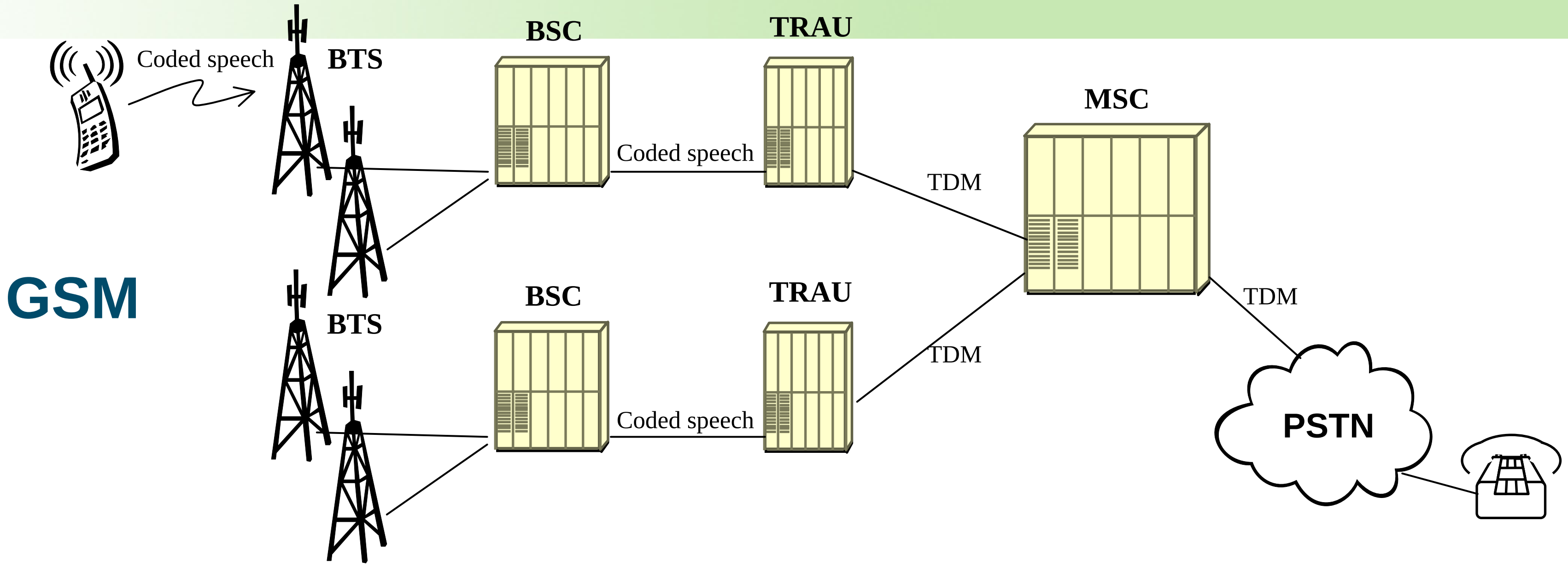
Introduction

- VQE techniques usually operate in the waveform domain.
- In GSM/UMTS networks, the signal coming from the mobile terminals has to be decoded, enhanced and encoded again.
- These operations introduce delays and are particularly prone to adding further quantization noise.
- Furthermore, they do not exploit the information already present in a packet of coded speech.

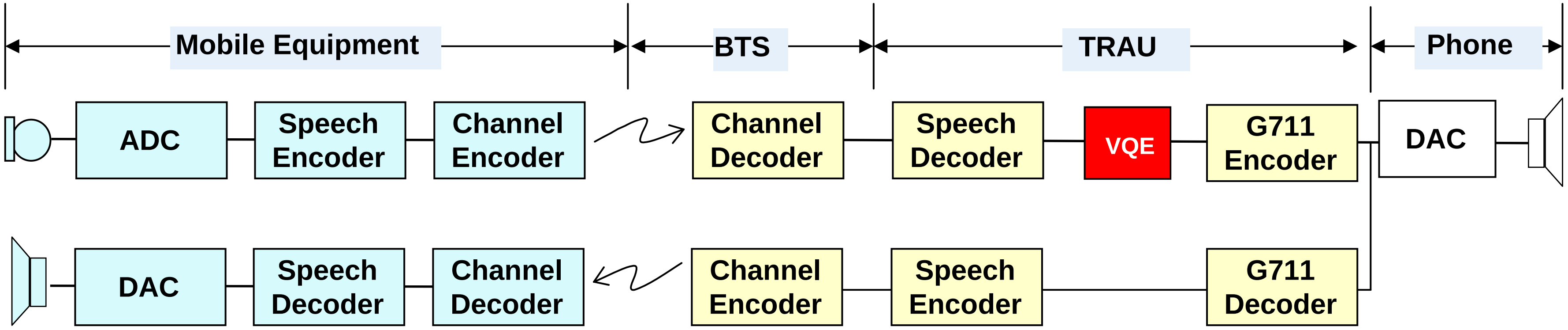
→ Solution: VQE in the coded domain

Voice Quality Enhancement

VQE processing location in the network



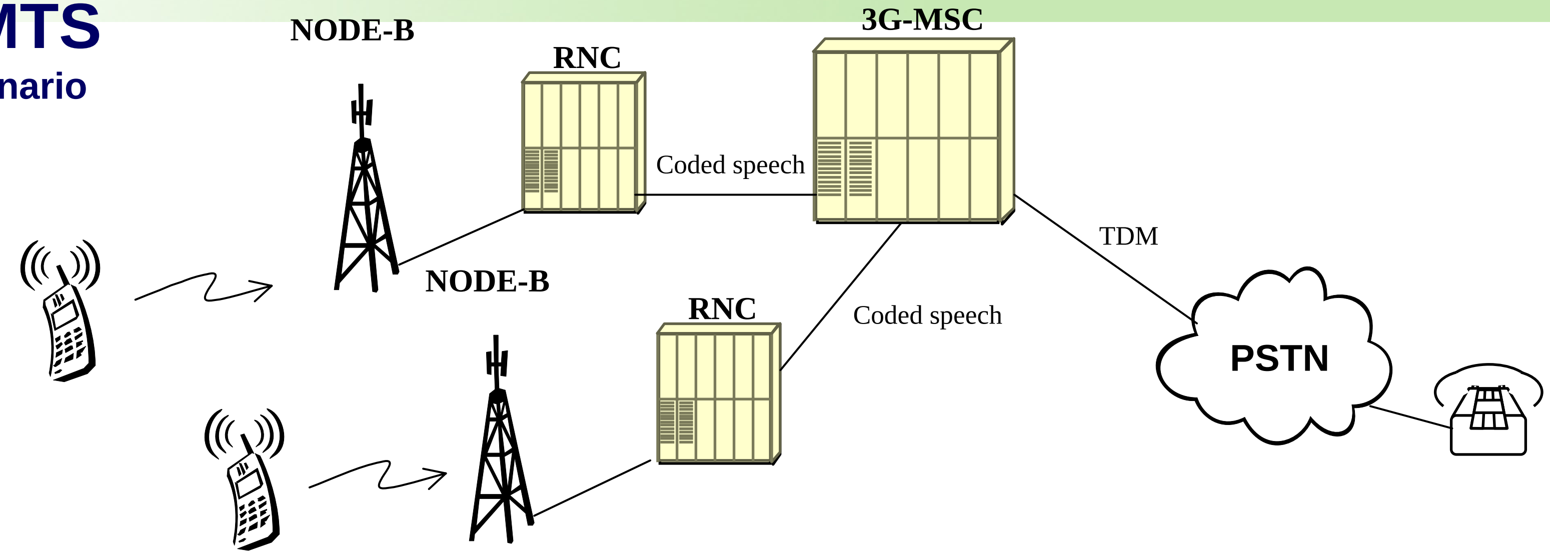
VQE performed on linear PCM samples after the speech decoder:



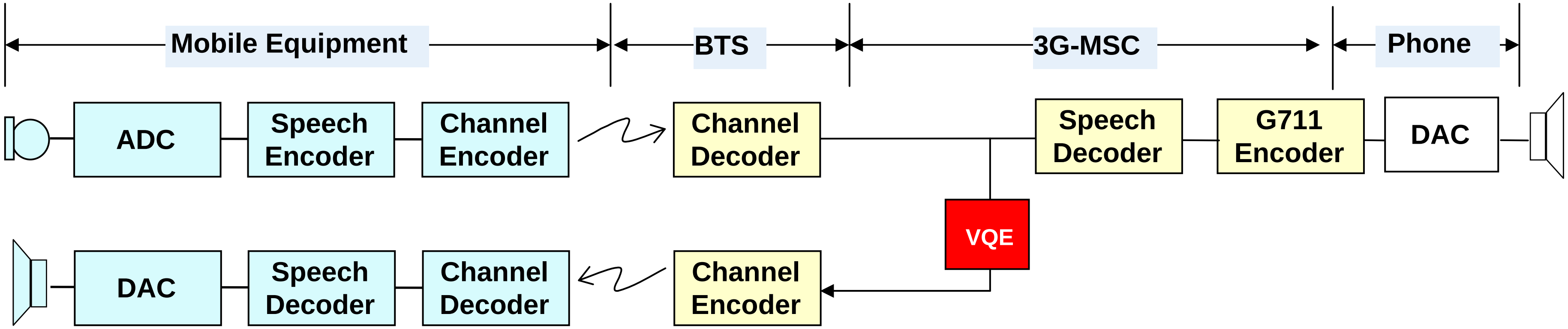
Voice Quality Enhancement

VQE processing location in the network – Next evolution

UMTS
scenario



Moving VQE before speech decoder or transcoder…



Thesis objectives

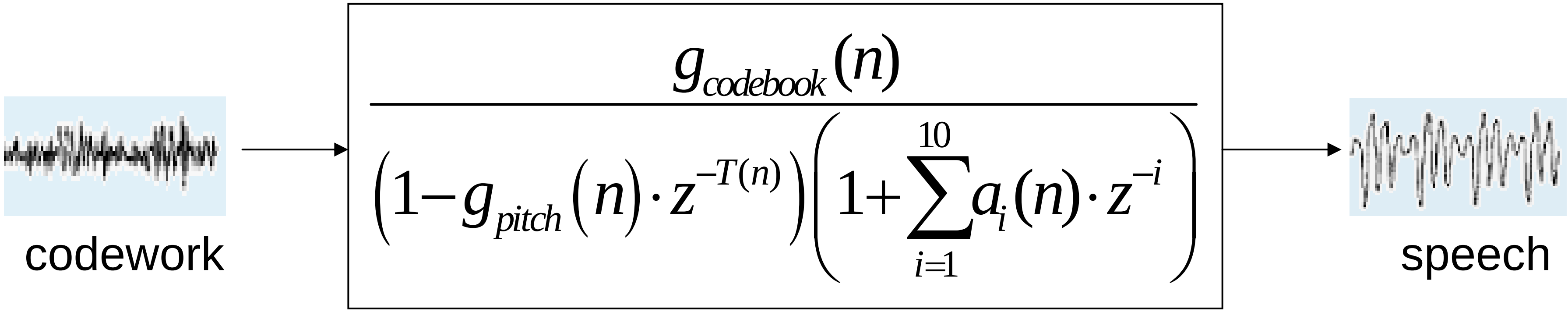
- Statistical analysis of the ACELP- AMR (*Adaptive Multi-Rate*) parameters

- *Voice Activity Detector* in the coded domain.
    - Performs the discrimination exploiting the statistical behavior of the set of parameters that characterize a segment of coded speech signal

- *Acoustic Echo Cancellation* in the coded domain.
    - Working directly on the coded parameters

Codec AMR 12.2 *kbit/s*

- Parameters where we work on:
  - 10 LPC coefficients
  - Pitch gain and lag (LTP order 1)
  - Fixed codebook gain



codework

$$\frac{g_{codebook}(n)}{\left(1-g_{pitch}(n)\cdot z^{-T(n)}\right)\left(1+\sum_{i=1}^{10} a_i(n)\cdot z^{-i}\right)}$$

speech

n-th subframe

Voice Activity Detection

- Discrimination between noise and voice AMR frames.
- Necessary for a good implementation of the VQE algorithms.

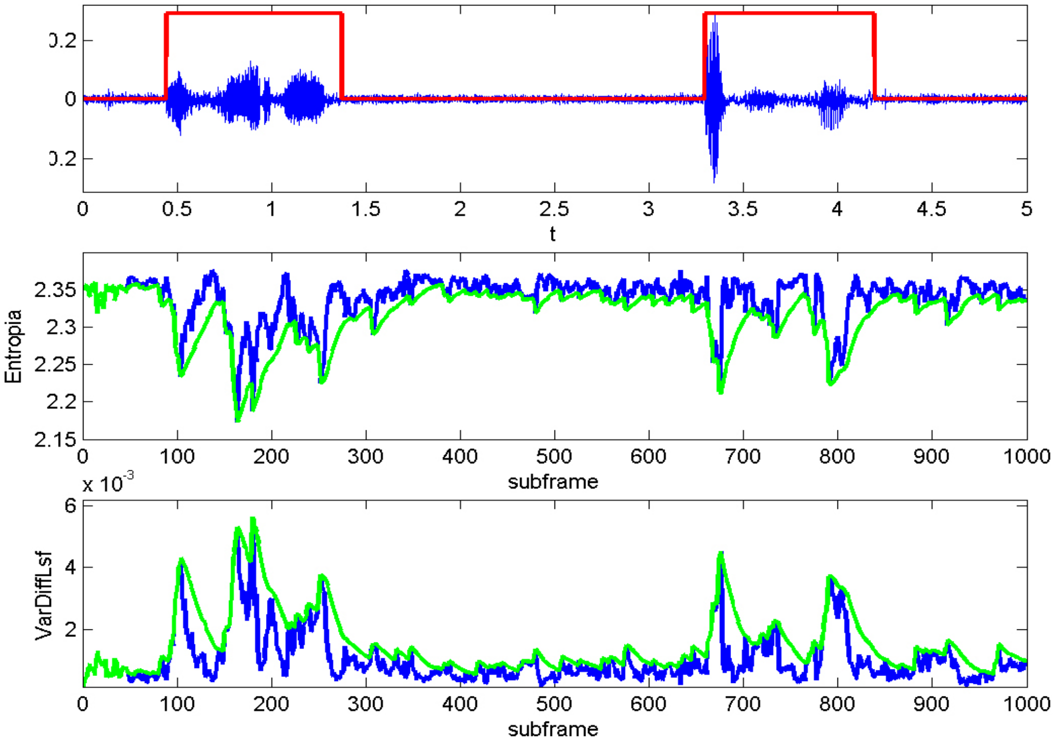
Voice Activity Detection

discriminative measures - LSFs

$$lsf' = (l_1, l_2 - l_1, l_3 - l_2, ..., l_{10} - l_9, \pi - l_{10})$$

$$Entropy = -\sum_{n=1}^{9}\left[\frac{\left|lsf'(n)\right|^2}{\sum_{n=1}^{9}\left|lsf'(n)\right|^2}\log_2\left(\frac{\left|lsf'(n)\right|^2}{\sum_{n=1}^{9}\left|lsf'(n)\right|^2}\right)\right]$$

$$VarDiffLsf = -\sum_{n=1}^{9}\left[lsf'(n) - \frac{1}{9}\sum_{n=1}^{9}lsf'(n)\right]^2$$

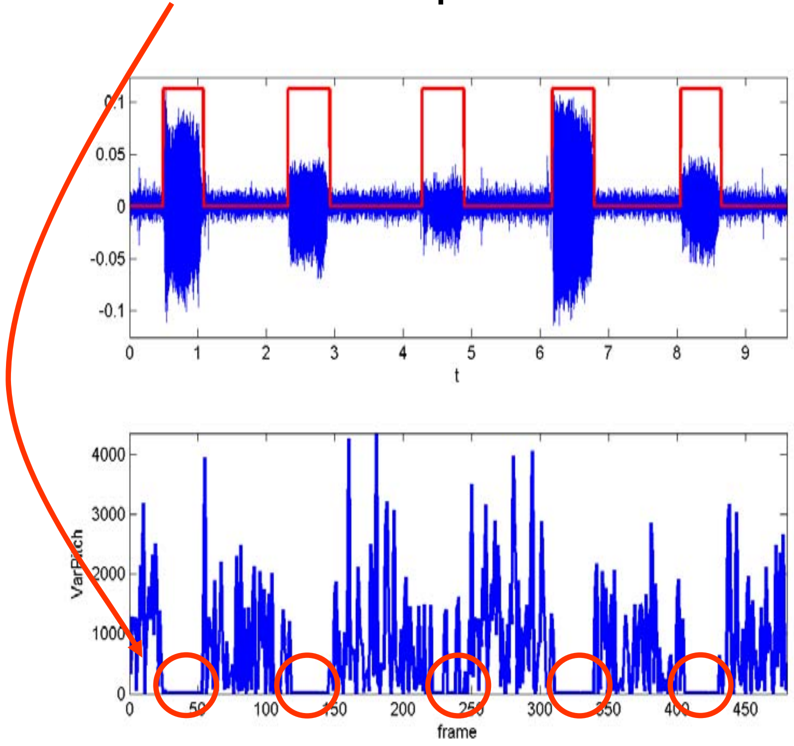


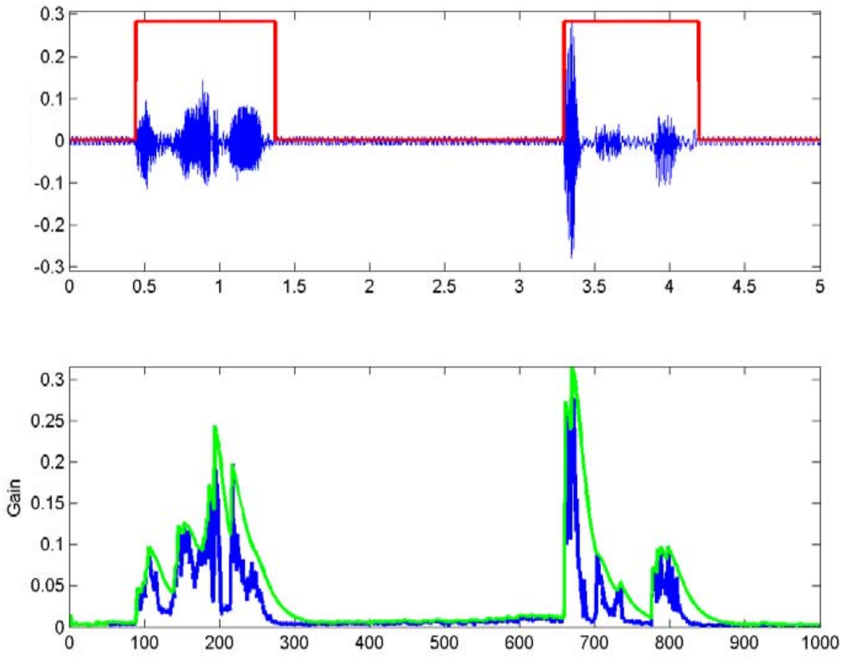
Voice Activity Detection

discriminative measures - pitch lag and gain

- pitch lag remains constant during vocalized speech

- algebraic codebook gain directly related to the energy



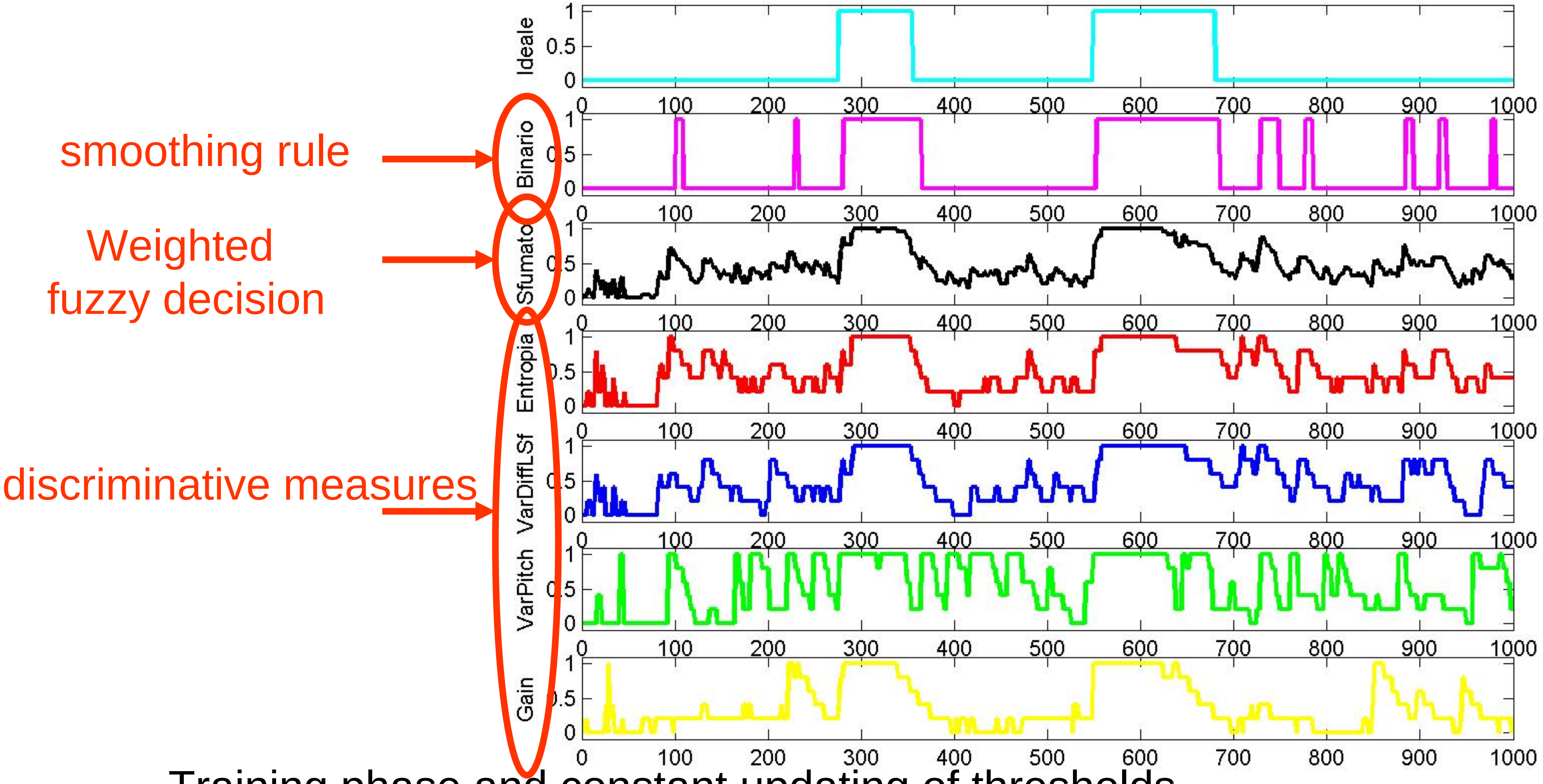


$$\mathrm{var}Pitch = \sum_{n=1}^{4}\left[T_0(n) - \frac{1}{4}\sum_{n=1}^{4}T_0(n)\right]^2$$

$$Gcodebook = Gcodebook$$

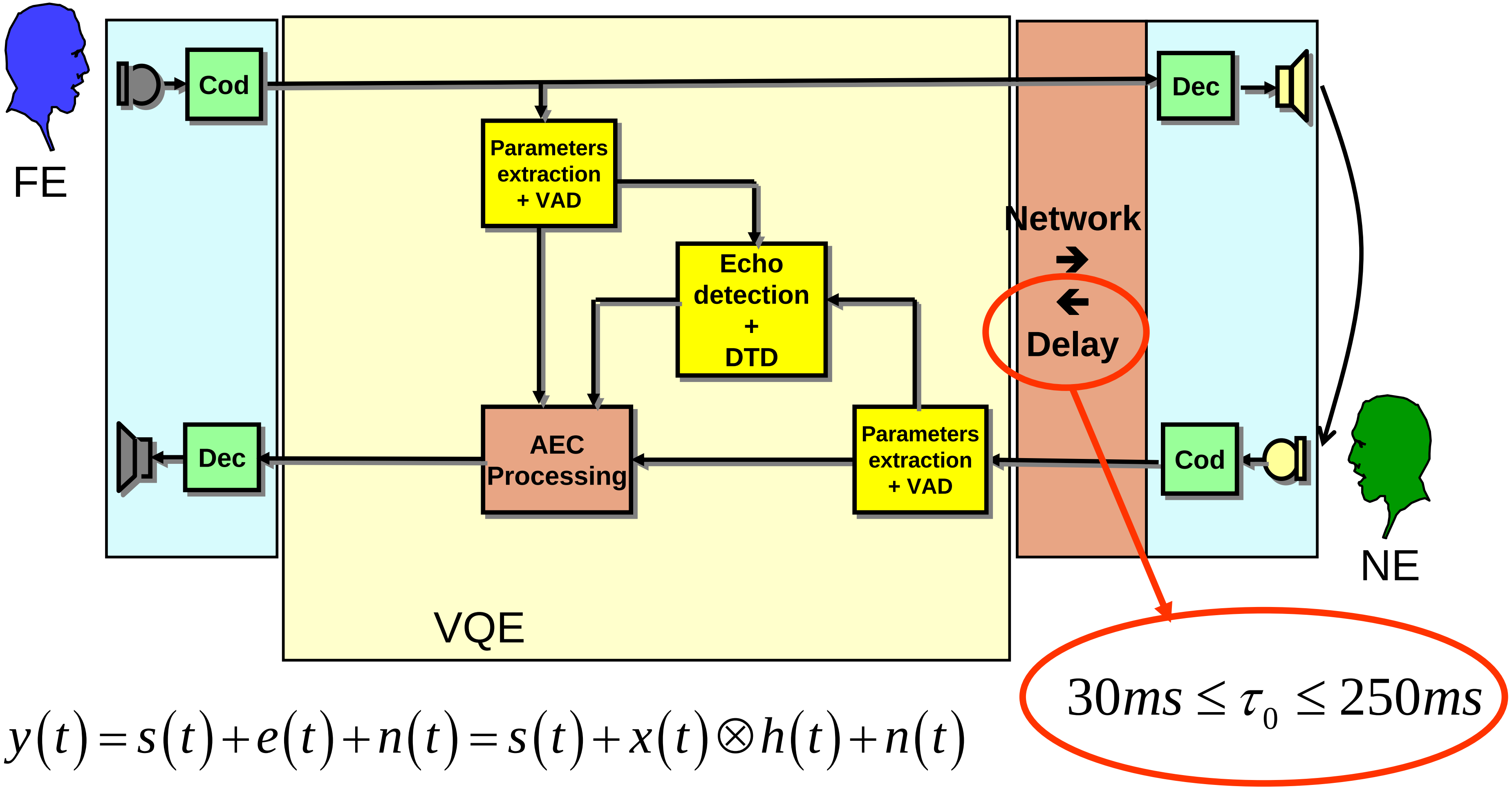
Voice Activity Detection

example



→ Training phase and constant updating of thresholds. Robustness to change in environments

Acoustic Echo Cancellation



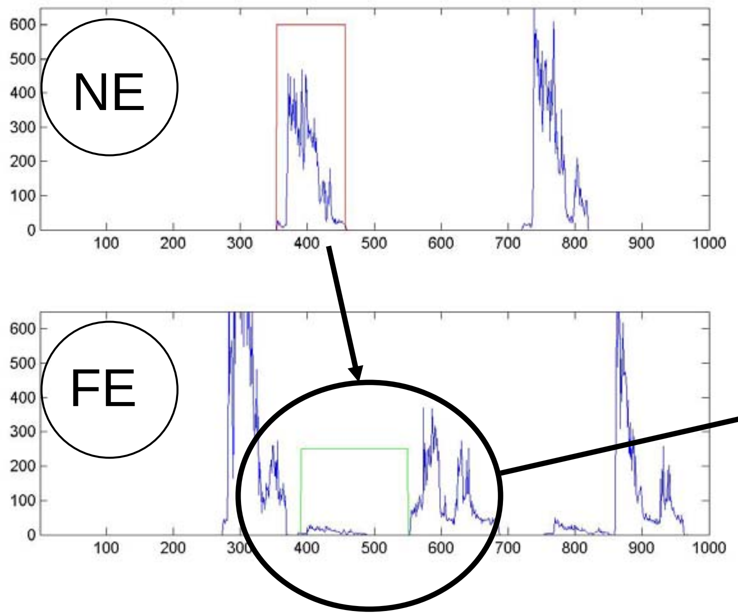
$$30ms \leq \tau_0 \leq 250ms$$

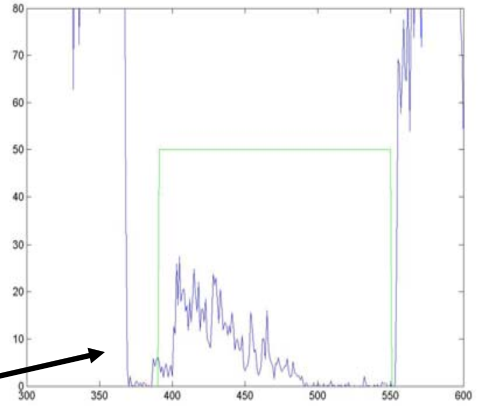
$$y(t) = s(t) + e(t) + n(t) = s(t) + x(t) \otimes h(t) + n(t)$$

$$\longrightarrow h(t) = \alpha \cdot \delta(t - \tau_0)$$

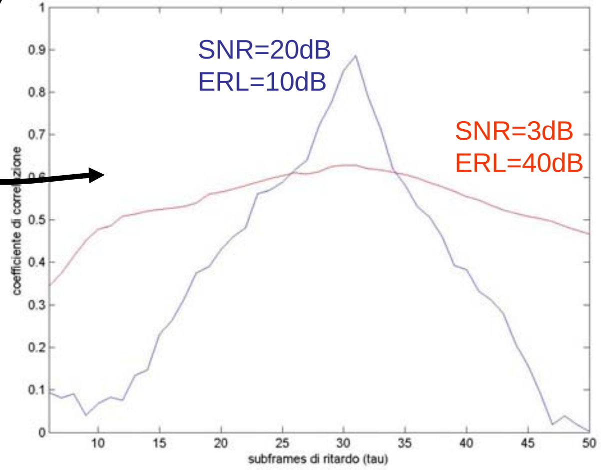
Acoustic Echo Cancellation

echo detector: initial estimate of network delay





$$\hat{\tau}_0 = \arg\max_{\tau} r_{xy}(\tau)$$



$$r_{xy}(\tau) = \frac{E\left[\left(x(n+\tau)-\mu_x\right)\left(y(n)-\mu_y\right)\right]}{\sqrt{E\left[\left(x(n+\tau)-\mu_x\right)^2\right]E\left[\left(y(n)-\mu_y\right)^2\right]}}, \quad \tau = 6,\ldots,50$$

$$r_{xy}(\tau) = \frac{\sum_{i=1}^{10} r_{lsf_x i, lsf_y i}(\tau) + r_{T_x, T_y}(\tau) + r_{g_{pitch,x}, g_{pitch,y}}(\tau) + r_{g_{fixed,x}, g_{fixed,y}}(\tau)}{13}$$

Acoustic Echo Cancellation

echo detector: updating network delay estimate

- Once the time near-end and far-end axis are aligned, we update iteratively the delay estimate, when:

$$VAD_x\left(n-\hat{\tau}_0\right)=1 \quad \wedge \quad VAD_y\left(n\right)=1$$

- cross-correlation calculation: $cc_{xy}\left(n,\tau\right)=\frac{E\left[x\left(m\right)y\left(m+\tau\right)\right]}{\sqrt{E\left[x^2\left(m\right)\right]E\left[y^2\left(m\right)\right]}}, \quad \tau=-20,\ldots,20$

$$cc_{xy}\left(n,\tau\right)=\frac{\sum_{i=1}^{10} r_{lsf_x i,lsf_y i}\left(\tau\right)+r_{T_x,T_y}\left(\tau\right)+r_{g_{pitch,x},g_{pitch,y}}\left(\tau\right)+r_{g_{fixed,x},g_{fixed,y}}\left(\tau\right)}{13}$$

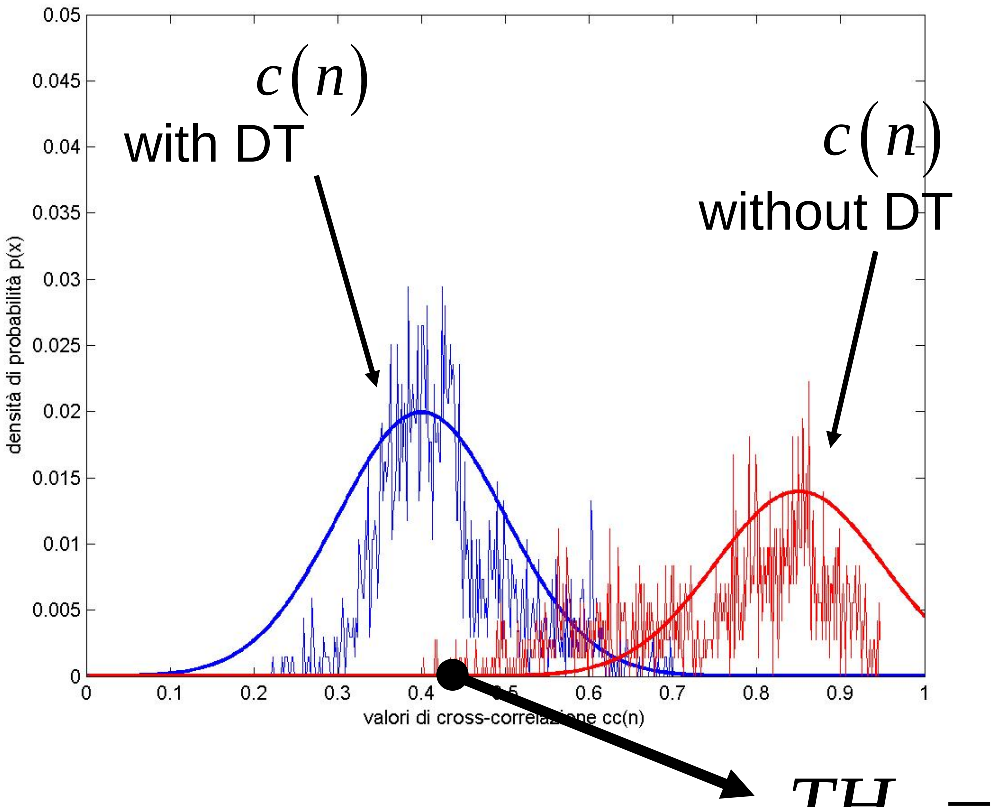
- define the maximum: $c\left(n\right)=\max cc_{xy}\left(n,\tau\right)$

- If: $c\left(n\right)>0.85$ then: $\delta\hat{\tau}_0\left(n\right)=\arg\max_{\tau} cc_{xy}\left(n,\tau\right) \longrightarrow$ update

$\longrightarrow$ $c\left(n\right)$ *echo likelihood* parameter. useful also for DTD and for the cancellation algorithms.

Acoustic Echo Cancellation

double-talk detector



$c(n)$ with DT

$c(n)$ without DT

$TH_c = 0.42$

$$P(D) \ll P(\bar{D})$$

$$P(D) = 0.05$$

Acoustic Echo Cancellation

AEC on the gains

$$H(z,n)=\frac{g_{fixed}(n)}{\left(1-g_{pitch}(n)\cdot z^{-T(n)}\right)\left(1+\sum_{i=1}^{10}a_i(n)\cdot z^{-i}\right)}$$

Fundamental to reduce the Energy level

Hyp: $g_y(n)=f\left(g_e(n),g_v(n),g_{bn}(n)\right)\approx g_e(n)+g_v(n)+g_{bn}(n)$

→ Normalized Least Mean Square

$$\underline{\hat{h}}(n+1)=\underline{\hat{h}}(n)+1.5\cdot c(n)\frac{\left(g_y(n)-\hat{g}_e(n)\right)}{\underline{g}_x{}^T(n)\underline{g}_x(n)}\underline{g}_x(n)$$

$$g_u(n)=g_y(n)-\hat{g}_e(n)=g_y(n)-\underline{\hat{h}}^T(n)\underline{g}_x(n)$$

Acoustic Echo Cancellation

AEC on the pitch lag

The spectral behaviour is still similar to the original echo signal

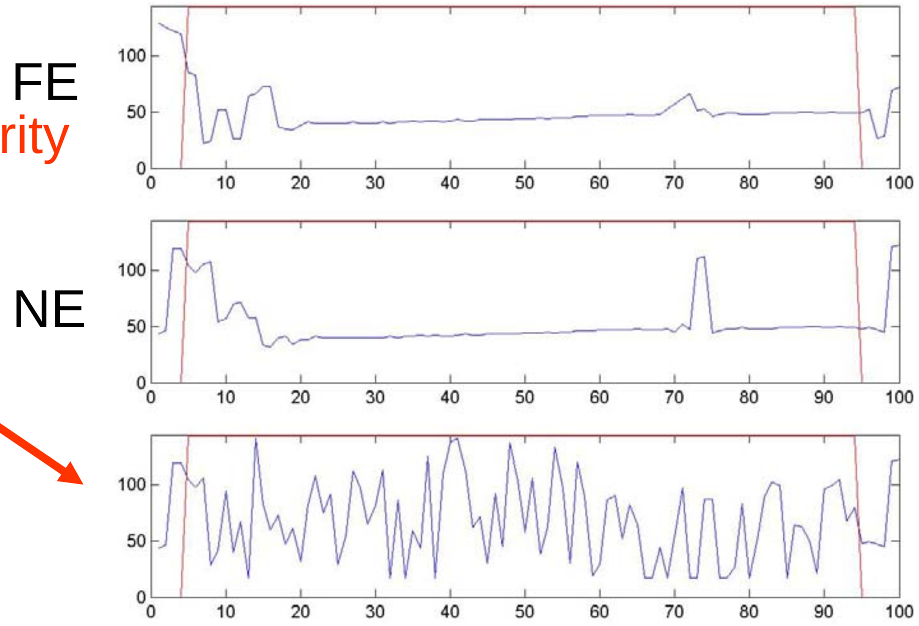
→ Solution: modify the spectrum of the AMR subframe

Pitch lag is basically constant during voiced speech

→ Solution: “break” this regularity

$$T_{pitch,u}(i) = T_r$$

$$\Omega_{T_r} = \{17,18,19,....,142,143\}$$

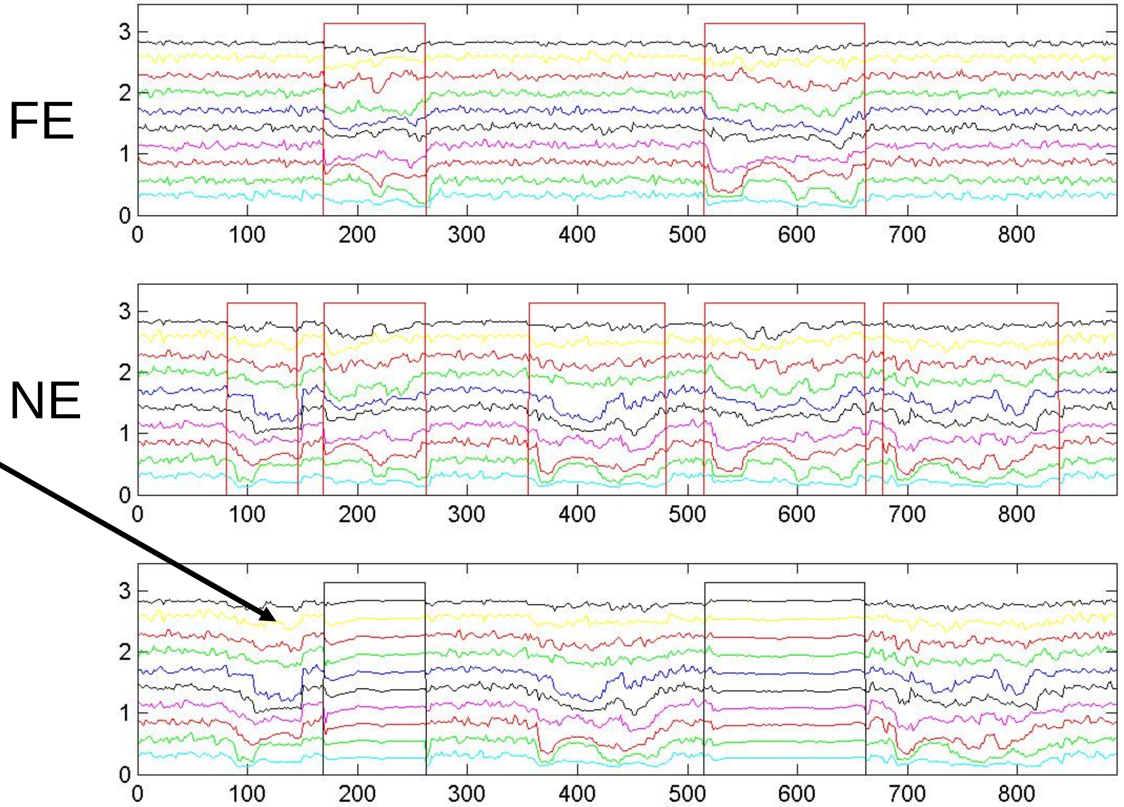


Acoustic Echo Cancellation

AEC on the LSFs

Modifying the LSFs, we can change the spectral coeherence of the signal

White noise example:
equidistant LSFs

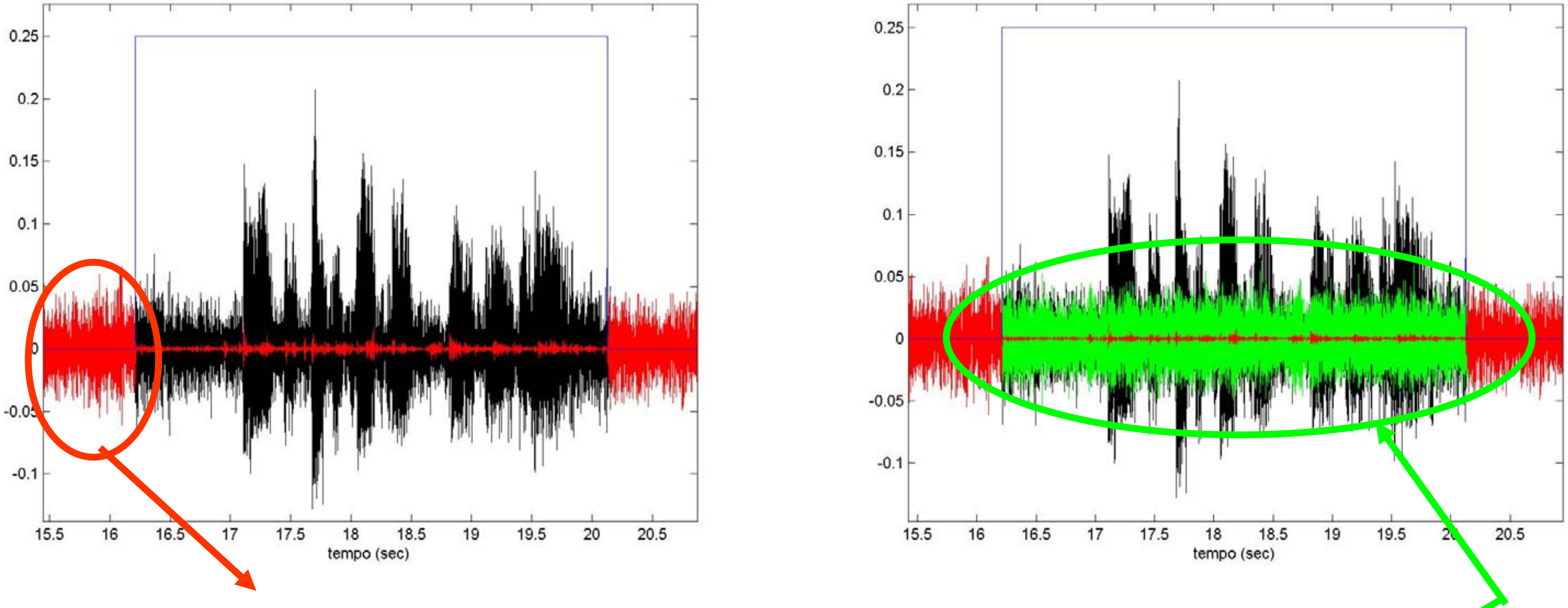


Noisy LSFs vector
updated when VAD=0

$$\hat{l}_{u,i}(n) = c(n) \cdot l_{bnoise}(n) + \left(1 - c(n)\right) \cdot l_i(n)$$

Acoustic Echo Cancellation

noise injection



Estimate noise statistics in the coded domain when VAD=0,
HMM-based noisy parameters generator

Conclusions

- VAD techniques have shown to be robust also with low SNR, with an accurate training phase.
- AEC techniques showed interesting performances, also comparable to linear-domain algorithms (20-25 dB ERLE) in average working conditions (Standard ITU)
  - PRO: possibility of modyfing the spectrum very easily
  - CON: difficult to increase ERLE in bad SNR and ERL conditions (not too relevant though)

→ Possible development of the algorithms in the near future