
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...

Mobile Equipment  ➔  BTS  ➔  3G-MSC  ➔  Phone

ADC ➔ Speech Encoder ➔ Channel Encoder ➔ Channel Decoder ➔ Speech Decoder ➔ G711 Encoder ➔ DAC

DAC ➔ Speech Decoder ➔ Channel Decoder ➔ Channel Encoder ➔ VQE ➔
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

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

\text{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

\[ lsf' = (l_1, l_2 - l_1, l_3 - l_2, \ldots, l_{10} - l_9, \pi - l_{10}) \]

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

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

discriminative measures - pitch lag and gain

- pitch lag remains constant during vocalized speech
- algebraic codebook gain directly related to the energy

\[
\text{varPitch} = \sum_{n=1}^{4} \left[ T_0(n) - \frac{1}{4} \sum_{n=1}^{4} T_0(n) \right]^2
\]

\[
G\text{codebook} = G\text{codebook}
\]
Voice Activity Detection

Example

- Weighted fuzzy decision
- Discriminative measures
- Training phase and constant updating of thresholds. Robustness to change in environments.
Acoustic Echo Cancellation

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

\[ h(t) = \alpha \cdot \delta(t - \tau_0) \]

\[ 30ms \leq \tau_0 \leq 250ms \]
Acoustic Echo Cancellation

echo detector: initial estimate of network delay

\[ \hat{r}_0 = \arg \max_{\tau} r_{xy}(\tau) \]

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

\[
r_{xy}(\tau) = \frac{\sum_{i=1}^{10} r_{ls\tau,ls\tau}(\tau) + r_{x\tau, y\tau}(\tau) + r_{\gamma_{pitch,x}, \gamma_{pitch,y}}(\tau) + r_{g_{fixed,x} g_{fixed,y}}(\tau)}{13}
\]
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:

  \[ VAD_x(n - \hat{\tau}_0) = 1 \land VAD_y(n) = 1 \]

- Cross-correlation calculation:

  \[
  cc_{xy}(n, \tau) = \frac{E[x(m)y(m+\tau)]}{\sqrt{E[x^2(m)]E[y^2(m)]}}, \quad \tau = -20, \ldots, 20
  \]

  \[
  cc_{xy}(n, \tau) = \frac{\sum_{i=1}^{10} r_{lsf,i,lsf,i}(\tau) + r_{T_x,T_y}(\tau) + r_{g_{pitch,x} \cdot g_{pitch,y}}(\tau) + r_{g_{fixed,x} \cdot g_{fixed,y}}(\tau)}{13}
  \]

- Define the maximum:

  \[
  c(n) = \max cc_{xy}(n, \tau)
  \]

- If: \( c(n) > 0.85 \) then:

  \[
  \delta \hat{\tau}_0(n) = \arg \max_{\tau} cc_{xy}(n, \tau) \quad \text{update}
  \]

\[ c(n) \quad \text{echo likelihood parameter. useful also for DTD and for the cancellation algorithms.} \]
Acoustic Echo Cancellation

double-talk detector

\[ P(D) \ll P(\bar{D}) \]
\[ P(D) = 0.05 \]
\[ TH_c = 0.42 \]
Acoustic Echo Cancellation
*AEC on the gains*

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

Fundamental to reduce the Energy level

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

Normalized Least Mean Square

\[
\hat{h}(n+1) = \hat{h}(n) + 1.5 \cdot c(n) \frac{g_y(n) - \hat{g}_e(n)}{g_x^T(n) g_x(n)} g_x(n)
\]

\[
g_u(n) = g_y(n) - \hat{g}_e(n) = g_y(n) - \hat{h}^T(n) g_x(n)
\]
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

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

$$\Omega_{T_r} = \{17, 18, 19, \ldots, 142, 143\}$$
Modifying the LSFs, we can change the spectral coherency of the signal.

\[ \hat{l}_{u,i}(n) = c(n) \cdot l_{b\text{noise}}(n) + (1 - c(n)) \cdot l_i(n) \]
Acoustic Echo Cancellation

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 modifying 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