DATA EMBEDDING IN SPEECH SIGNALS USING PERCEPTUAL MASKING

Ariel Sagi and David Malah

Technion - Israel Institute of Technology, Department of Electrical Engineering

Signal and Image Processing Laboratory

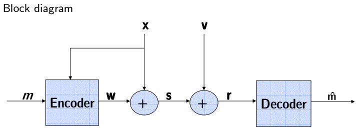
ABSTRACT

- **Data embedding technique** for **speech signals**, exploiting the **masking property** of the human auditory system
- Data embedding is performed by modifying the **Discrete Hartley Transform** (DHT) coefficients according to the principles of the **Scalar Costa Scheme** (SCS)
- **Data embedding parameters** of each subband are computed from the **auditory masking threshold function** and a **channel noise estimate**
- A **maximum likelihood detector** is employed in the decoder for **embedding parameters detection**
- The proposed technique is simulated by **embedding-data** in a **speech signal** transmitted over a **telephone line**

-1-

DATA-EMBEDDING MODEL

- Block diagram



- Notations

m message
$\mathbf{w}$ watermark signal
$\mathbf{x}$ host signal
$\mathbf{s}$ combined signal
$\mathbf{v}$ noise
$\mathbf{r}$ received signal
$\hat{m}$ decoded message

- Definitions

$$\mathrm{WNR} = 10\log_{10}\left(\frac{\sigma_w^2}{\sigma_v^2}\right) \text{ [dB]}$$
$$\mathrm{SWR} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_w^2}\right) \text{ [dB]}$$

-2-

DATA-EMBEDDING SCHEMES

- **Ideal Costa Scheme**
  Costa, 1983: "Writing on Dirty Paper", proved that for IID Gaussian host signal and IID Gaussian noise host signal interference can be completely avoided

$$C_{\mathrm{ICS}} = \frac{1}{2}\log_2\left(1+\frac{\sigma_w^2}{\sigma_v^2}\right)$$

- **Scalar Costa Scheme**
  Eggers & Girod suggested a **suboptimal practical** embedding rule, that uses **dithered uniform scalar quantizers**
  - Encode message m in $\mathbf{d} = d_1, d_2, \ldots, d_n$, where $d \in \{0, 1, \ldots, D-1\}$
  - Embed $\mathbf{d} = d_1, d_2, \ldots, d_n$ in $\mathbf{x} = x_1, x_2, \ldots, x_n$

$$s_n = (1-\alpha)x_n + \alpha\left(Q_\Delta\left\{x_n - \Delta\left(\frac{d_n}{D}\right)\right\} + \Delta\left(\frac{d_n}{D}\right)\right)$$

-3-

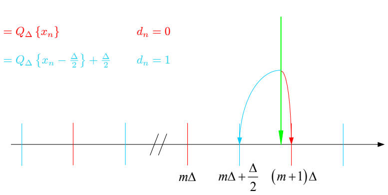
SCALAR COSTA SCHEME ENCODER

Example: $\{\alpha = 1, D = 2\}$

$$s_n = Q_\Delta\left\{x_n - \Delta\left(\frac{d_n}{2}\right)\right\} + \Delta\left(\frac{d_n}{2}\right)$$

$s_n = Q_\Delta\{x_n\}$ $\quad d_n = 0$

$s_n = Q_\Delta\{x_n - \frac{\Delta}{2}\} + \frac{\Delta}{2}$ $\quad d_n = 1$



-4-

SCALAR COSTA SCHEME DECODER

- The signal y_n is defined by

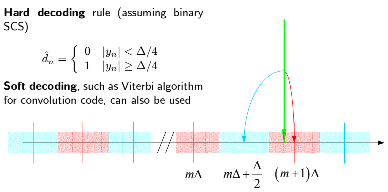
$$y_n = Q_\Delta\{r_n\} - r_n$$

and therefore $|y_n| \le \Delta/2$

- **Hard decoding** rule (assuming binary SCS)

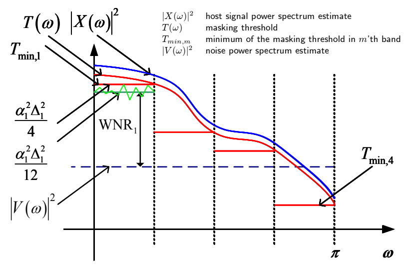
$$\hat{d}_n = \begin{cases} 0 & |y_n| < \Delta/4 \\ 1 & |y_n| \ge \Delta/4 \end{cases}$$

- **Soft decoding**, such as Viterbi algorithm for convolution code, can also be used



-5-

DATA-EMBEDDING USING PERCEPTUAL MASKING



-6-

SUBBAND PARAMETER DETERMINATION

- The subband average embedding-distortion can be expressed by

$$\sigma_{w,m}^2 = \frac{\alpha_m^2\Delta_m^2}{12} = \frac{10^{T_{min,m}/10}}{3}$$

- **Scale factor determination**
  Given a model or estimation of the subband noise variance $\sigma_{v,m}^2$, the scale factor α_m is given by

$$\alpha_m = \sqrt{\frac{\sigma_{w,m}^2}{\sigma_{w,m}^2 + 2.71\sigma_{v,m}^2}}$$

- **Quantization-step determination**
  The subband quantization-step value is given by

$$\Delta_m = \frac{2}{\alpha_m}10^{T_{min,m}/20}$$

- To improve the robustness and computational complexity, Δ_m is quantized, in the log domain, to one of $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$

-7-

DATA-EMBEDDING DOMAIN

- **Discrete Cosine Transform**

  The masking threshold function should be transformed to the DCT domain

- **Discrete Fourier Transform**

  The DFT is a complex valued transform

- **Discrete Hartley Transform**

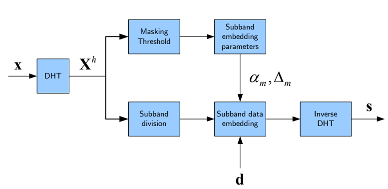
$$X_k^h = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} x_n \mathrm{cas}\left(\frac{2\pi}{N}nk\right); \quad k = 0, 1, \ldots, N-1$$

where $\mathrm{cas}(x) \triangleq \cos(x) + \sin(x)$

-8-

ENCODER STRUCTURE

- Block diagram



-9-

DECODER STUCTURE

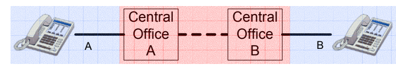
The decoder comprises of:

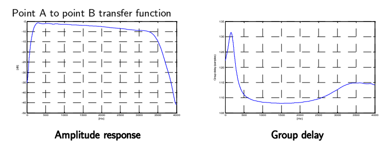
- **Adaptive equalizer** which reduces the channel spectral distortion
- Joint subband **embedded-data presence detection** and **quantization-step determination**
- **Embedded-data decoding**

-10-

TELEPHONE CHANNEL MODEL

- Amplitude and phase distortion, u-law or A-law quantization noise, Circuit (white Gaussian) noise





-11-

CHANNEL EQUALIZATION

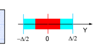
- Common adaptive equalization algorithms
  - Time domain: NLMS, RLS
  - Frequency domain
- There is need for a training sequence for the above algorithms. In case of a telephone conversation, listening to the training sequence can be annoying
  - Solution: Select a chosen **audio/speech signal** as a **training sequence**
- Blind equalization algorithms
  - Pros: A training sequence is not needed
  - Cons: Not practical in our scenario, where data is embedded in a much stronger host signal.

-12-

QUANTIZATION-STEP DETERMINATION (1/2)

- The estimated quantization-step will be one of $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$
- For each subband the decoder decides on the tested quantization-step values, and defines $\mathbb{G}$ as their indexes
- For each quantization-step index of $\mathbb{G}$, the decoder calculates the subband demodulated DHT coefficients

$$Y_{m,k}^g = Q_{\Delta^g}\{R_{m,k}\} - R_{m,k}; \quad g \in \mathbb{G}$$



where m is the subband index and k is the coefficient index

- Define two possible hypotheses
  - H_0: correct quantization-step, with PDF $p(Y|H_0)$
  - H_1: incorrect quantization-step, with PDF $p(Y|H_1)$

-13-

QUANTIZATION-STEP DETERMINATION (2/2)

- The log-likelihood ratio (LLR), for each quantization-step index of $\mathbb{G}$

$$L_m^g = \log\left\{\frac{p(\mathbf{Y}_m^g|H_0)}{p(\mathbf{Y}_m^g|H_1)}\right\}; \quad g \in \mathbb{G}$$

- The quantization-step index that maximize the LLRs, L_m^g

$$g^* = \arg\max_{g\in\mathbb{G}} L_m^g$$

- The estimated quantization step in the m'th subband is the quantization-step value that maximize the LLR

$$\hat{\Delta}_m = \Delta^{g^*}$$

- The maximal LLR, denoted by $L_m^{g^*}$, is used in the subband embedded-data presence detection rule

$$\mathbb{I}_m = \begin{cases} 1; & L_m^{g^*} > T \\ 0; & L_m^{g^*} \le T \end{cases}$$

-14-

PERFORMANCE EVALUATION

Simulation setup

- Telephone channel model, RLS equalization (256 taps). Embedding data in frames with speech presence (detected by energy VAD).

The proposed data-embedding system performance is evaluated by the following objectives:

- Transparency MOS=3.95
- Embedding-rate RATE=(8000/256)*16*0.6=300[bits/sec.]
  (1 subband/frame, 16 information bits/subband, VAD detection rate=0.6)
- Robustness BER(coded)=~$1\cdot10^{-4}$

System	MOS	VAD rate	Error correction code	Embedding -rate (bps)	BER (coded)
Original, adaptively selecting 1 subband/frame	3.95	0.6	BCH (16,31)	300	~$1\cdot10^{-4}$
Updated results, adaptively selecting 2 subband/frame	3.9	0.8	Golay (12,23)	600	~$7\cdot10^{-5}$

-15-