DATA EMBEDDING IN SPEECH SIGNALS USING PERCEPTUAL MASKING

Ariel Sagi and David Malah

Technion - Israel Institute of Technology
Department of Electrical Engineering, Haifa 32000, Israel.
e-mail: sagiaril@tx.technion.ac.il, malah@ee.technion.ac.il

ABSTRACT

In this paper, a data embedding technique for speech signals, exploiting the masking property of the human auditory system, is presented. The signal in the frequency domain is partitioned into subbands. The data embedding parameters of each subband are computed from the auditory masking threshold function and a channel noise estimate. Data embedding is performed by modifying the Discrete Hartley Transform (DHT) coefficients according to the principles of the Scalar Costa Scheme (SCS). A maximum likelihood detector is employed in the decoder for embedded-data presence detection and data-embedding quantization-step estimation. We demonstrate the proposed data embedding technique by simulation of data embedding in a speech signal transmitted over a telephone line. The demonstrated system achieves transparent data-embedding at the rate of 300 information bits/second with a bit-error-rate of approximately 10^{-4}. The proposed technique outperforms spread spectrum (SS) based data-embedding techniques for speech signals.

1. INTRODUCTION

A data embedding (also known as data hiding or digital watermarking) system should fulfill the following requirements. It should embed information *transparently*, meaning that the quality of the host signal is not degraded perceptually by the presence of embedded data. It should be *robust*, meaning that the embedded data could be decoded from the watermarked signal, even if it is distorted or attacked. The data embedding *rate* is also of importance in some applications.

Eggers and Girod [4], motivated by Costa's work [3], proposed a practical data embedding scheme. The scheme uses uniform scalar quantization in the data embedding process. The capacity of SCS is typically higher then other proposed schemes (for example, schemes based on SS [2][10] or quantization index modulation (QIM) [1]). However, the general method in [4] does not take into consideration human perception models, such as human visual or human auditory models.

Many SS-based data embedding techniques do use a perceptual model in the embedding process [10]. However, the disadvantage of these techniques is their low embedded data rate, which is a consequence of the SS principles.

In this paper we combine the general principles of SCS with an auditory perceptual model, obtaining a data-embedding system for speech signals.

The paper is organized as follows. In section 2 we review the main principles of SCS. In sections 3,4 we present the data embedding encoder and decoder structure, respectively. We describe simulation results in section 5, followed by conclusions in section 6.

2. SCS PRINCIPLES

A general model for data communication by data embedding can be described as follows: The binary representation of a message m, denoted by a sequence $\mathbf{b}$, is encoded into a coded sequence $\mathbf{d}$ (by a forward error-correction channel-coder such as block codes or convolutional codes). The encoder embeds the coded data $\mathbf{d}$ into the host signal $\mathbf{x}$ producing the transmitted signal $\mathbf{s}$. A deliberate or an undeliberate attack might modify the signal $\mathbf{s}$ into a distorted signal $\mathbf{r}$ and impair data transmission. The decoder aims to detect the embedded data from the received signal $\mathbf{r}$. In blind data-embedding systems, the host signal $\mathbf{x}$ is not available at the decoder.

In this paper $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{r}$ denote the *representation* vectors of the original, transmitted, and received signal, respectively. Lower case x_n, s_n and r_n refer to their respective n'th element. Representation vector elements could be time samples, frequency coefficients or any other representation elements of the corresponding signal.

2.1 Data embedding

According to SCS, the transmitted signal elements are additively composed of the host signal and the watermark signal, i.e., $s_n = x_n + w_n$. The watermark signal elements are given by $w_n = \alpha q_n$, where α is a scale factor and q_n is the host signal element quantized according to the data d_n:

$$q_n = Q_\Delta\left\{x_n - \Delta\left(\frac{d_n}{D} + k_n\right)\right\} - \left(x_n - \Delta\left(\frac{d_n}{D} + k_n\right)\right). \quad (1)$$

In this paper we assume a binary SCS, i.e., a SCS with an alphabet size of $D = 2$, that is, $d_n \in \mathbb{D} = \{0,1\}$. $Q_\Delta\{\cdot\}$ denotes scalar uniform quantization with step-size Δ, and $k_n \in [0,1)$ denote the elements of a cryptographically secure pseudo-random sequence $\mathbf{k}$. For simplicity, we assume in the following that the sequence $\mathbf{k}$ is not in use, i.e. $k_n \equiv 0$. The noise elements are given by $v_n = r_n - s_n$, and the watermark-to-noise ratio (WNR) is defined as:

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

where σ_w^2,σ_v^2 are the variances of the watermark and noise signals elements, respectively. SCS embedding depends on two parameters: the quantizer step-size Δ and the scale factor α. For a given watermark power σ_w^2 , these two parameters are related via:

$$\sigma_w^2 = \frac{\alpha^2\Delta^2}{12}. \quad (3)$$

Eggers and Girod found experimentally an approximate analytical expression for the optimum value of α, in the sense of maximizing the capacity of SCS, which is [4]:

$$\alpha_{\text{SCS,approx}} = \sqrt{\frac{\sigma_w^2}{\sigma_w^2 + 2.71\sigma_v^2}}. \quad (4)$$

(3) and (4) lead to $\Delta_{\text{SCS,approx}} = \sqrt{12(\sigma_w^2 + 2.71\sigma_v^2)}$. In the case of $\alpha = 1$, SCS and QIM scheme [1] have the same embedding rule.

2.2 Data extraction

Data detection is applied to a signal $\mathbf{y}$, whose elements are computed from the received signal elements r_n by:

$$y_n = Q_\Delta\{r_n\} - r_n. \quad (5)$$

For binary SCS, $|y_n| \le \Delta/2$, and for proper detection, y_n should be close to zero if $d_n = 0$ was embedded, and close to $\pm\Delta/2$ if $d_n = 1$. A hard decoding rule assigns:

$$\hat{d}_n = \begin{cases} 0 & |y_n| < \Delta/4 \\ 1 & |y_n| \ge \Delta/4 \end{cases}. \quad (6)$$

Soft-input decoding algorithms, e.g., a Viterbi decoder like the one used for decoding convolutional codes, can be used here too to decode the most likely transmitted message $\hat{m}$, from the signal $\mathbf{y}$.

3. DATA-EMBEDDING ENCODER

In this section, the data-embedding encoder structure is explained. The proposed encoder performs data embedding in the frequency domain, in separate subbands, by utilizing a masking threshold function (MTF).

3.1 Computation of subband masking thresholds

The MTF is an estimate of the masking threshold at each spectral location. It determines the maximum allowed distortion in each band by the embedded watermark, so as to keep transparency to the human listener. Under the constraint of transparent embedding, the proper utilization of the MTF will result in the maximal possible WNR.

The computation of the MTF is based on MPEG's psycho-acoustic model [6]. The standard supports several common sampling frequencies of audio signals. Some modifications in the masking model implementation should be made in the case of speech signals sampled at 8KHz.

The speech signal, $\mathbf{x}$, is divided into non-overlapping frames of length N. The l-th frame is denoted by $\mathbf{x}^l$ with elements $\{x_p^l = x_{lN+p};\quad 0 \le p \le N-1\}$. The MTF, $\{S_k;\quad 0 \le k \le N/2\}$, with k denoting a discrete frequency index, is calculated for each frame. The positive frequency band is divided into M equal width subbands ($M < N/2$). The subband masking threshold (SMT) in each subband is set to the minimum of the MTF value in that subband:

$$S_{min,m} = \min_{k \in m\text{'th subband}} S_k; \quad m = 1,2,\ldots,M \quad (7)$$

3.2 Data embedding domain

For each type of a host signal there is a need to decide on the appropriate embedding domain. The use of a frequency domain auditory masking model naturally leads to the choice of the frequency domain representation of a signal as the embedding domain. In other words, the frequency domain coefficients of the host signal are modified according to (1).

For reasons explained below, we use the DHT coefficients for data embedding. The DHT, X_k^l, of the signal frame $\mathbf{x}^l$, is defined by [9]:

$$X_k^l = \frac{1}{\sqrt{N}} \sum_{p=0}^{N-1} x_p^l \text{cas}\left(\frac{2\pi}{N} pk\right); \quad k = 0,\ldots,N-1, \quad (8)$$

where $\text{cas}(x) \triangleq \cos(x) + \sin(x)$. As for the DFT, the transform elements are periodic in k with period N. We prefer the DHT over other common frequency domain representations, such as the discrete Fourier transform (DFT) or the discrete cosine transform (DCT).

The reason for preferring the DHT over the DFT is because the latter is a complex transform, while the DHT is a real one, and there are fast algorithms for the computation of the DHT [9], similar to the those used for the computation of the DFT.

The DFT is commonly used for computing the MTF [6]. Yet, the need for complex arithmetic can be completely eliminated by using the direct relation between the DFT and DHT, given by:

$$\text{Re}\{F_k\} = \frac{1}{2}[X_{N-k} + X_k]; \quad \text{Im}\{F_k\} = \frac{1}{2}[X_{N-k} - X_k], \quad (9)$$

where X_k and F_k denote the DHT and DFT of a signal $\mathbf{x}$, respectively. Therefore, in the proposed scheme the DHT is calculated to obtain a representation of the signal for data embedding, followed by the computation of the DFT components by (9), that are then used to compute the MTF. Although the DCT is also a real transform, it does not provide the same simplicity in computing the MTF as the DHT.

The computed DHT coefficients are divided into M equal-width subbands, like the MTF, and data embedding in the coefficients of each subband is carried out with the subband SCS parameters $\{\alpha_m, \Delta_m\}$.

3.3 Determination of subband SCS parameters

The *maximal* embedding distortion in (1) is $\alpha^2\Delta^2/4$, while the average embedding distortion is $\alpha^2\Delta^2/12$. Distortion in the m'th subband that is greater than $S_{min,m}$ (7) may be audible. It is required therefore that the subband maximal embedding-distortion will be bounded from above by the SMT. By equating the subband maximal embedding-distortion with the SMT: $10\log_{10}[\alpha_m^2\Delta_m^2/4] = S_{min,m}[\text{dB}]$, the subband average embedding-distortion can be expressed in terms of $S_{min,m}$ by:

$$\sigma_{w,m}^2 = \frac{\alpha_m^2\Delta_m^2}{12} = \frac{10^{S_{min,m}/10}}{3}. \quad (10)$$

Assuming that a channel-noise model is given, and denoting the noise variance estimate in the m'th subband by $\sigma_{v,m}^2$, the value of the subband scale factor, α_m, can be obtained from (10) and (4).

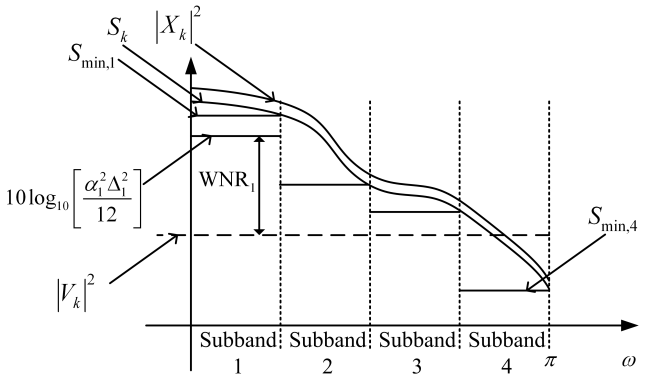

Figure 1: A schematic drawing of a speech signal power spectrum estimate, $|X_k|^2$, divided into 4 subbands; MTF - S_k; The SMT - $S_{min,m}$, is marked by the horizontal solid lines. Additive white gaussian noise (AWGN) source power spectrum estimate $|V_k|^2$ is marked by the dashed line. The WNR in the first subband is also marked.

Formally, the subband quantization-step value is given now from (10), by $\Delta_m^* = \frac{2}{\alpha_m} 10^{S_{min,m}/20}$. However, to improve the robustness of the quantization-step estimation in the decoder, the applied subband quantization-step is selected to be one of a finite pre-defined set of quantization-step values, denoted by $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$, that will be known also at the decoder. Therefore, the quantization-step in the m'th subband is obtained by quantizing, in the log domain, the above computed Δ_m^*, yielding:

$$\Delta_m = 10^{\frac{D_m}{20}}, \tag{11}$$

where $D_m \triangleq c \left\lfloor \frac{S_{min,m}+20\log_{10}[2/\alpha_m]}{c} \right\rfloor$ and the constant c is the quantization step of Δ_m^* in [dB].

3.4 Selecting subbands for data embedding

There can be various approaches for selecting subbands for data embedding. A possible criterion is to embed data in a specific subband only if speech is present and the estimated WNR in that subband is greater then a given threshold value, that is set according to a target BER value. Another criterion is to embed data in a predefined fixed number of subbands, chosen dynamically from the set of all subbands, as those that provide the maximal estimated WNR, when speech is present.

If it is decided to embed data in the m'th subband, the DHT coefficients are modified according to the SCS embedding rule shown in (1), with the parameters $\{\alpha_m, \Delta_m\}$. The modified DHT coefficients are inverse transformed to obtain the transmitted signal.

Figure 1 schematically demonstrates the proposed data embedding procedure.

4. DATA-EMBEDDING DECODER

In this section, the data-embedding decoder structure is explained. First, in subsection 4.1, a channel equalizer, compensating for the channel filtering effect, is described. Detection of embedded-data presence in each subband is needed (subsection 4.2) when the encoder chooses dynamically the subbands for data embedding. When the decoder detects embedded-data presence, it estimates the quantization-step used at the encoder. Finally, with the estimated quantization step, the embedded data can be decoded.

4.1 Channel equalization

A common model of the channel (or attack) is an AWGN source. We assume here a more complex channel model, modelled by a linear time–invariant (LTI) filter followed by an AWGN source. Such a channel is typical in many applications (e.g., a telephone channel) and may adversely affect the system's performance. To compensate for the channel filtering effect, we apply adaptive equalization. There is a variety of adaptive equalization algorithms in the literature [5], such as the LMS and RLS algorithms. A training signal is used in the training stage. Blind equalization algorithms are used for equalizing data communication channels, but to the knowledge of the authors there is no blind equalization algorithm that performs well in our scenario, where data is embedded in a much stronger host signal. The signal filtered by the trained equalizer is of degraded quality because of the equalization noise enhancement. Therefore, this signal is used only for decoding the embedded data. Following channel equalization, the decoder performs frame synchronization as well.

4.2 Combined embedded-data presence detection and quantization-step estimation

To test for embedded-data presence in a given subband, we define two possible hypothesis:

- Hypothesis H_0: embedded-data is present in the given subband.
- Hypothesis H_1: embedded-data is absent in the given subband.

We define two probability density functions $p(\mathbf{Y}_m|H_0)$ and $p(\mathbf{Y}_m|H_1)$. $\mathbf{Y}_m$ is the result of (5) applied to the DHT coefficients of the received signal in the m'th subband, denoted by $\mathbf{R}_m$. In [4] the functions $p(\mathbf{Y}_m|H_0)$ and $p(\mathbf{Y}_m|H_1)$ are denoted by $p(\mathbf{y}|H_0)$ and $p(\mathbf{y}|H_1)$, and their properties are explained.

The decoder computes the quantization step, $\Delta_{m,\mathrm{dec}}$, by repeating the same steps done in the encoder for computing the quantization step (7),(10)-(11). The estimated quantization-step is one of the set of quantization steps, $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$. Afterwards, a test set of quantization-step indices is chosen, denoted by $\mathbb{G}$. For example, if $\{\Delta_{m,\mathrm{dec}} = \Delta^i; \quad i \in \{0,1,\ldots,J-1\}\}$, we can choose the test set as $\mathbb{G} = \{i-1, i, i+1\}$. Using the test set $\mathbb{G}$, (5) is applied to the received subband DHT coefficients $\mathbf{R}_m$, to obtain $\{\mathbf{Y}_m^g; \quad g \in \mathbb{G}\}$.

The log-likelihood ratio (LLR), for each quantization step of the test set, is computed from:

$$L_m^g = \log\left[\frac{p(\mathbf{Y}_m^g|H_0)}{p(\mathbf{Y}_m^g|H_1)}\right] = \log\left[\frac{\prod_{k=0}^{N/M-1} p(Y_{m,k}^g|H_0)}{\prod_{k=0}^{N/M-1} p(Y_{m,k}^g|H_1)}\right]; \quad g \in \mathbb{G}. \tag{12}$$

The LLR, L_m^g, is a measure of the validity of the assumption that Δ^g is the quantization step used in the encoder.

The embedded-data presence detector first computes the quantization-step index that maximize the LLRs, L_m^g:

$$g^* = \arg\max_{g \in \mathbb{G}} L_m^g. \tag{13}$$

The maximal LLR from (13), denoted by $L_m^{g^*}$, is then used in the subband embedded-data presence detection rule:

$$\mathbb{I}_m = \begin{cases} 1; & L_m^{g^*} > T \\ 0; & L_m^{g^*} \leq T \end{cases}, \quad (14)$$

where T is a decision threshold. The detector decides that embedded-data is present in the m'th subband if $\mathbb{I}_m = 1$, and that it is absent if $\mathbb{I}_m = 0$.

If $\mathbb{I}_m = 1$, the quantization step in the m'th subband is estimated by the quantization-step value that maximize the LLR according to (13), i.e.,

$$\hat{\Delta}_m = \Delta^{g^*}. \quad (15)$$

Using the estimated subband quantization step, $\hat{\Delta}_m$, the embedded data can be extracted using the procedure described in subsection 2.2.

5. SIMULATION RESULTS

We demonstrate the proposed data embedding technique by embedding data in a speech signal transmitted over a telephone line.

The telephone line causes amplitude and phase distortion combined with PCM quantization noise and AWGN. The telephone line simulation model is based on a ITU-T standard [7]. The full band is partitioned into $M = 8$ subbands. The first and last subbands are not used for data embedding because the telephone line has a large attenuation in the frequency ranges of 0-300Hz and 3400-4000Hz. To equalize the telephone line we used a RLS equalizer with 256 taps.

Data embedding transparency was evaluated by the perceptual evaluation of speech quality (PESQ) tool [8]. The evaluation results are equivalent to a mean opinion score (MOS) scale of [0-4.5]. A score between 3.6 and 4.2 is widely accepted as relating to good quality. The comparison is between the original speech signal and the received signal before equalization. The encoder parameters were empirically set by constraining the average MOS result to be more then 3.9. Informal listening test confirmed the PESQ tool evaluation.

For channel coding we used Bose-Chaudhuri-Hoequenghem (BCH) error-correction coding with 16 information bits and 31 coded bits per subband. Since there are 32 coefficients in each subband, an extra parity bit is concatenated to the coded bits, resulting in 32 coded bits per subband. Data is embedded in single subband in each speech frame when an energy based voice activity detector (VAD) detects speech presence. The selected subband is chosen to be the one with the maximal estimated subband WNR. The average number of frames for which the VAD detected speech was approximately 60 percent of the total number of frames. The data-embedding rate when speech is present is equal to 500 information bits/second, corresponding to an average data transmission rate of 300 information bits/second. This system configuration is denoted as system A.

The results are compared to an embedding system, denoted as system B, operating without the use of a MTF. As for system A, a single subband is selected for data embedding, when the VAD detects speech presence. The selected subband is chosen to be the one with the maximal subband variance. For the computation of the embedding parameters, $\{\alpha_m, \Delta_m\}$, we replace $S_{min,m}$ (7) with $\{10\log_{10}[\sigma_{x,m}^2] - \rho\}$[dB], where $\sigma_{x,m}^2$ denotes the m'th subband coefficients variance, and the constant ρ is chosen such that the degradation in speech quality, compared on a MOS scale, will be approximately equal to that of system A.

The MOS (as measured by PESQ) is approximately 3.95 for both systems, and the data-embedding rate is equal to 300 information bits/second. The BER values obtained are approximately 10^{-4} and $1.2 \cdot 10^{-2}$, for systems A and B, respectively. The subband average WNR values, in the subbands where embedded-data presence was detected, are approximately 21dB and 12dB, respectively.

6. CONCLUSIONS

We have presented a data-embedding algorithm in speech signals that applies an auditory perceptual masking model to a SCS-based data embedding system, and outperforms SS based data-embedding techniques. We propose methods for computing subband data-embedding parameters in the encoder, and for estimating the subband embedded-data presence and quantization step in the decoder. Although we demonstrated the proposed data-embedding technique for speech signals, the technique can also be used, with slight modifications, for data embedding in audio signals.

REFERENCES

[1] B. Chen and G. W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 44(4):1423–1443, May 2001.

[2] Q. Cheng and J. Sorensen. Spread spectrum signaling for speech watermarking. volume 3, pages 1337–1340. ICASSP, 2001.

[3] M. H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[4] J. J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod. Scalar costa scheme for information embedding. *IEEE Transactions on Signal Processing*, 51(4):1003–1019, April 2003.

[5] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 3'rd edition, 1996.

[6] ISO/IEC. Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s–part 3: audio. Technical Report IS 11172-3, 1992.

[7] ITU-T. Netwrok transmission model for evaluating modem performance over 2-wire voice grade connections. Technical Report V.56 bis, August 1995.

[8] ITU-T. Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Technical Report P.862, February 2001.

[9] H. Sorensen, D. Jones, C. Burrus, and M. Heideman. On computing the discrete hartley transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(5):1231–1238, October 1985.

[10] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney. Robust audio watermarking using perceptual masking. *Signal Processing*, 66(3):337–355, 1998.