CCIT Report #603

October 2006

Bandwidth Extension of Telephone-Speech Aided by Data Embedding

Ariel Sagi* and David Malah

Technion - Israel Institute of Technology

Department of Electrical Engineering, Haifa 32000, Israel

Email: ariels@il.ibm.com, malah@ee.technion.ac.il

Abstract

In this paper a system for bandwidth extension of telephone-speech, aided by data embedding, is presented. The proposed system uses the transmitted analog narrowband speech signal as a carrier of the side information needed to carry out the bandwidth extension. The upper-band of the wideband speech is reconstructed at the receiving end from two components: a synthetic wideband excitation signal, generated from the narrowband telephone-speech and a wideband spectral envelope, parametrically represented and transmitted as embedded-data in the telephone-speech. We propose a novel data embedding scheme, in which the scalar Costa scheme is combined with an auditory masking model allowing high rate transparent embedding, while maintaining a low bit error rate. The signal is transformed to the frequency domain via the discrete Hartley transform (DHT) and is partitioned into subbands. Data is embedded in an adaptively chosen subset of subbands by modifying the DHT coefficients. In our simulations, high quality wideband speech was obtained from speech transmitted over a telephone line (characterized by spectral magnitude distortion, dispersion, and noise), in which side information data is transparently embedded at the rate of 600 information bits/second and with a bit error rate of approximately $3 \cdot 10^{-4}$. In a listening test, the reconstructed wideband speech was preferred (at different degrees) over conventional telephone speech in 92.5% of the test utterances.

Keywords and phrases: speech bandwidth extension, auditory masking, data embedding, digital watermarking, scalar Costa scheme.

*Currently at IBM Haifa Research Laboratory, Haifa 31905, Israel.

1 Introduction

Public telephone systems reduce the bandwidth of the transmitted speech signal from an effective frequency range of 50Hz to 7KHz to the range of 300Hz to 3.4KHz. The reduced bandwidth leads to a characteristic thin and muffled sound of the so called *telephone speech*. Listening tests have shown that the speech bandwidth affects the perceived speech quality [1]. Artificially extending the bandwidth of the *narrowband* (NB) speech signal can result in both higher intelligibility and higher subjective quality of the reconstructed *wideband* (WB) speech. Usually, the information required for speech bandwidth extension (SBE)[2] is generated from the received NB speech or transmitted separately. Typically, the latter method results in higher quality of the reconstructed WB speech.

A unique SBE system in which the transmission from and to the talker's handset is analog, and hence particulary suitable for the public telephone system, is suggested in this paper. The proposed scheme uses the speech signal as a carrier of the side information required for SBE, by auditory-transparent data-embedding, eliminating the need of an additional channel for the side information while providing high quality reconstructed WB speech. This SBE application could be attractive for enhancement of the conventional public telephone system, requiring only DSP hardware operating at the receive and transmit sides of the telephone connection.

The structure of the SBE system is shown in Figure 1. The input to the system is a WB speech signal, denoted by $\mathbf{s}_{\mathrm{WB}}$, which is fed in parallel into the SBE encoder and data-embedding blocks. The SBE encoder extracts the highband (HB) spectral parameters which are embedded in the telephone-band frequency range of the WB input signal (i.e., in the NB signal) by the data-embedding block. The modified NB speech is transmitted over a telephone-channel. At the receiver, adaptive equalization is applied to reduce the channel spectral distortion. The embedded-data is extracted from the NB speech signal at the channel equalizer output, and used by the SBE decoder to reconstruct WB speech, denoted by $\hat{\mathbf{s}}_{\mathrm{WB}}$.

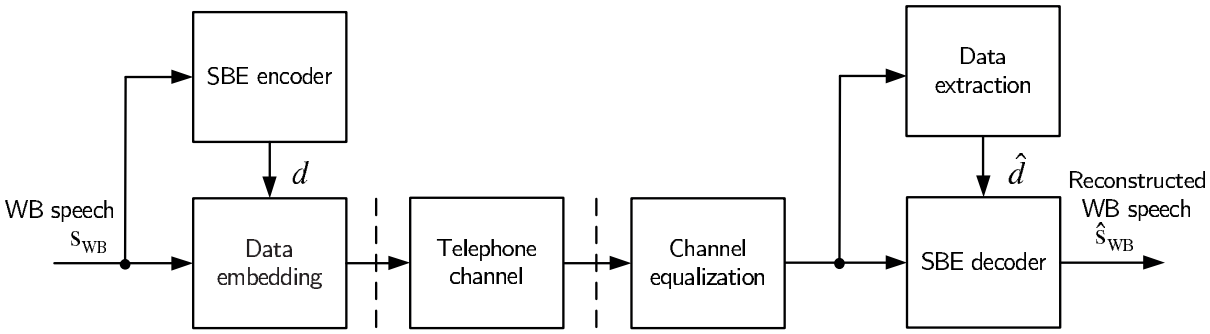

Figure 1: Speech bandwidth extension (SBE) system description.

The authors of [3], motivated by Costa's work [4], proposed a practical data embedding scheme, known as the scalar Costa scheme (SCS). The capacity of SCS is typically higher than other proposed schemes, e.g., schemes based on spread-spectrum (SS) [5][6] or quantization index modulation (QIM) [7]. However, the general method in [3] does not take into consideration human perception models, such as human visual or human auditory models. SS-based data embedding techniques that use a perceptual model in the embedding process were reported in [5][6]. However, the disadvantage of this techniques is low embedded data rate, which is a consequence of the SS principle. The authors of [8] proposed a data embedding scheme for speech, which is also a part of a SBE application. In the data-embedding encoder of [8] an excitation signal is first generated by filtering the NB speech signal with its corresponding linear prediction analysis filter to produce an excitation signal. Then, the excitation signal is projected to a subspace, where data embedding is applied using the vectorial form of QIM [7]. The NB speech with embedded data is produced by back projecting the modified subspace signal

to the excitation signal space, and then filtering the excitation signal with the corresponding linear prediction synthesis filter. The effect of the linear prediction analysis/synthesis filtering can be interpreted as noise shaping of the watermark signal which then follows the spectral characteristics of the speech. In the data-embedding decoder the identical transformation from the NB speech signal to the subspace signal is implemented, which follows data extraction.

In this paper we propose a novel combination of the SCS data-embedding method with an auditory masking model. In the proposed embedding scheme, the signal in the frequency domain is partitioned into subbands and the data embedding parameters for each adaptively selected subband are computed from the auditory masking threshold function and a channel noise estimate. An effective choice of the embedding domain, namely the *discrete Hartley transform* (DHT), is suggested and is found to have an advantage over the more common DCT and DFT domains. Data is embedded by modifying the DHT coefficients according to the principles of the SCS. A maximum likelihood detector is employed at the decoder for embedded-data presence detection and data-embedding quantization-step estimation. Partial details and preliminary results of the proposed data embedding scheme were reported by us in [9], without any consideration of the current application, i.e., speech bandwidth extension.

The telephone line causes amplitude and phase distortion combined with μ-law (or A-law) quantization noise and additive white Gaussian noise (AWGN). In [10][8] techniques for data-embedding in telephone-speech are proposed, but only the channel noise (PCM, μ-law, ADPCM, AWGN) is treated, disregarding the spectral distortion caused by the channel. In this work, we apply adaptive equalization to reduce the channel spectral distortion. Although the channel model in our work includes spectral distortion and dispersion, the achievable data-rate is much higher than the data-rate reported in [10][8]. For the AWGN channel model of [10], the achievable BER in our simulations is lower than the one reported in [10], and at the same time the achievable data-rate is much higher.

This paper is organized as follows. The SBE encoder and decoder structures are described in Section 2. In Section 3 the main principles of SCS are briefly reviewed and the combination of SCS with an auditory perceptual model is described. Results of subjective listening tests and objective evaluations are presented in Section 4, followed by conclusions in Section 5.

2 Speech Bandwidth Extension

In this section the part of the system performing SBE is described. We first describe the general principles of SBE systems in Section 2.1, and continue with the proposed SBE encoder and decoder structures details in Sections 2.2 and 2.3, respectively.

2.1 Principles of Speech Bandwidth Extension

Most of the works on SBE [11][12] use linear prediction (LP) techniques [13]. By these techniques, the WB speech generation at the receiving end is divided into two separate tasks. The first task is the generation of a WB excitation signal, and the second task is to determine the WB spectral envelope, represented by linear prediction coefficients (LPCs) or transformed versions like line spectral frequencies (LSF). Once these two components are generated, WB speech is regenerated by filtering the WB excitation signal with the WB linear prediction synthesis filter.

The generation of the WB excitation signal and the WB spectral envelope can be done by solely using the received NB speech signal [12][14]. The implicit assumption of such an approach is that there is correlation between the low and high frequencies of the speech signal. In [12], a dual codebook in which part of the codebook contains NB codewords and the other part contains highband (HB) codewords is proposed. A chosen NB codeword, which is the most similar to the input NB spectral envelope, points to a HB codebook. From this HB codebook, a HB codeword is chosen. In [14], a statistical approach based on a hidden Markov model is

used, which takes into account several features of the NB speech. Another approach is to code and transmit side information about the HB portion of the speech signal. The WB speech is then reconstructed at the encoder from the NB speech, and the received side information. This approach is hybrid, because it artificially regenerates the high-frequency excitation information from the NB speech signal, and obtains the high-frequency envelope information from the side information [8][15][16][17]. Some systems, e.g. [18], make use of both correlation between the low and high frequencies of the speech signal and side information, for the generation of the HB portion of the speech signal. The quality of WB speech generated by the hybrid approach is usually significantly better than the quality of WB speech generated by the NB speech only based approach.

In this work we use the hybrid approach, with the side information being embedded in the NB speech, like [8]. However, our proposed SBE and data-embedding schemes are different from the schemes suggested in [8].

2.2 SBE Encoder Structure

The SBE encoder extracts the HB spectral parameters that will be embedded in the NB speech signal. The parameters include a gain parameter and spectral envelope parameters for each frame of the original WB speech signal.

The structure of the SBE encoder is shown in Figure 2. The input to the SBE encoder is the original WB speech signal, denoted by $\mathbf{s}_{\mathrm{WB}}$. The WB speech signal is fed in parallel into three branches. We first describe the structure of each branch and in the sequel provide the details of the main blocks.

Upper Branch In this branch, the WB speech is passed through a 2:1 decimation system (composed of a low pass filter and a 2:1 down-sampler), obtaining a NB speech signal, denoted by $\mathbf{s}_{\mathrm{NB}}$. A time-domain LP analysis is performed on the NB signal, and the NB excitation (or residual) signal is obtained by inverse filtering the NB speech signal by the analysis filter. The NB excitation signal, denoted by $\mathbf{e}_{\mathrm{NB}}$ is then used for WB excitation regeneration at the encoder. The encoder reconstructed WB excitation signal is denoted by $\hat{\mathbf{e}}_{\mathrm{WB}}$.

Middle Branch In this branch, the WB signal is analyzed by applying, like [8], a selective LP analysis [19] to its HB, in the range 3-8KHz. The selective LP coefficients, $\mathbf{a}_{\mathrm{HB}}$, are converted into the LSF [20] representation, $\boldsymbol{\omega}_{\mathrm{HB}}$. The selective LSFs are quantized using a vector quantizer. The LSFs codebook index is one of the transmitted parameters via data embedding. The quantized selective LSFs are transformed into WB LPCs, denoted by $\hat{\mathbf{a}}_{\mathrm{WB}}$, which correspond to the reconstructed WB spectral envelope. For the purpose of determining an appropriate HB gain parameter, the WB LPCs are used to synthesize the WB reconstructed speech signal at the encoder, denoted by $\tilde{\mathbf{s}}_{\mathrm{WB}}$. In comparison, in [8] the selective LP coefficients are converted into the *cepstral* domain and are quantized by a vector quantizer.

Lower Branch In the lower branch, the HB gain parameter, denoted by g_{HB}, is computed by minimizing the spectral distance between the original and synthesized WB speech signals, in the 3-8KHz frequency range. After computing the gain, it is quantized, and the quantized gain index is transmitted.

The transmitted information in each analysis frame thus includes the LSF codebook index and the gain index (i.e., the indices of the parameters $\hat{\boldsymbol{\omega}}_{\mathrm{HB}}$ and $\hat{g}_{\mathrm{HB}}$, marked by dashed lines).

In the next subsections, the details of the main SBE encoder blocks are given.

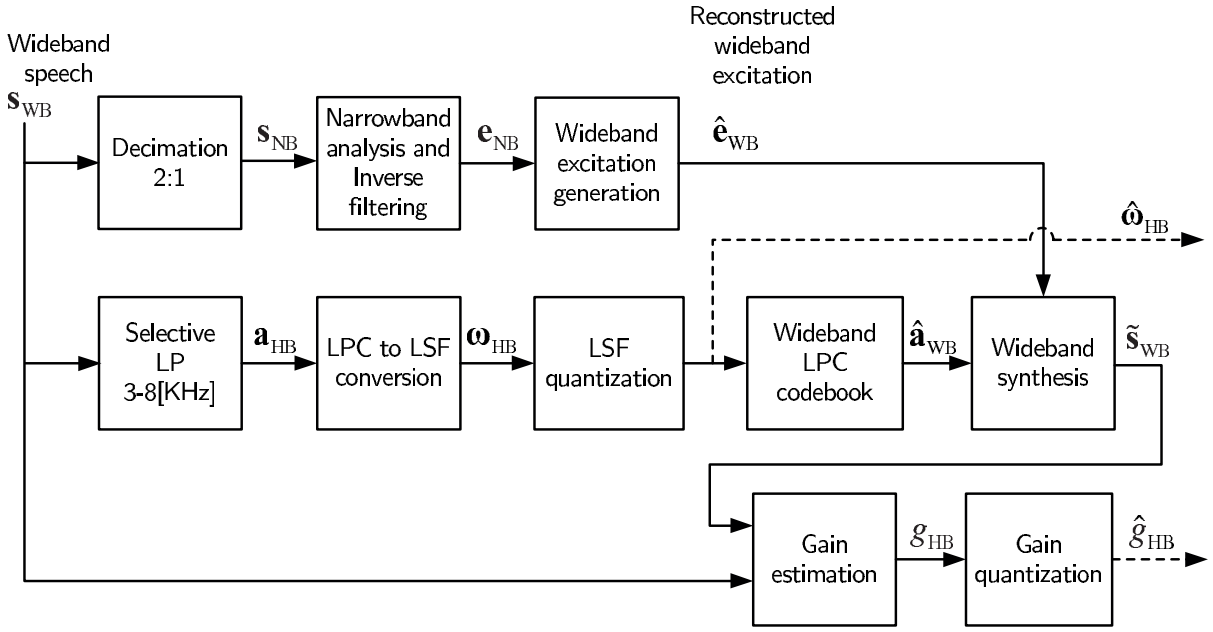

Figure 2: SBE Encoder structure.

2.2.1 Wideband Excitation Generation Block

The WB excitation can be artificially generated from the NB excitation signal by one of the methods described in [21]. The NB excitation signal is the output of inverse filtering by the LP analysis filter, applied to the NB speech signal. As shown in Figure 3, the NB excitation signal, $\mathbf{e}_{NB}$, is first passed through a 1:2 interpolation system (composed of a low pass filter and a 1:2 up-sampler) to the WB speech sampling rate. It is known that rectifiers and limiters typically expand the bandwidth of a signal. In our case, the interpolated NB excitation is passed through a full-wave rectifier, which performs sample by sample rectification [21]. The interpolated NB excitation is combined with the HB portion of the rectified signal, to produce an artificially extended WB excitation, denoted by $\tilde{\mathbf{e}}_{WB}$. This artificially extended WB excitation has a downward tilt in the high-frequencies due to the rectification operation. The tilt can be flattened by a whitening filter that performs inverse filtering. The filter is obtained by an LP analysis of the artificially extended WB excitation, $\tilde{\mathbf{e}}_{WB}$. The output of the whitening filter, which is the reconstructed WB excitation signal, is denoted by $\hat{\mathbf{e}}_{WB}$.

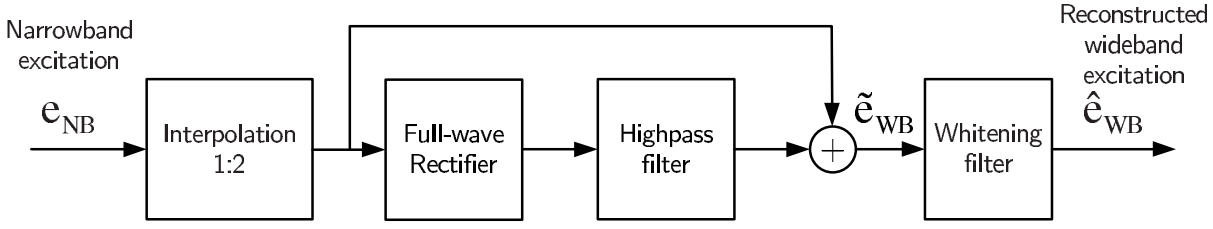

Figure 3: Artificial WB excitation generation.

2.2.2 Selective LP, LPC to LSF conversion, and LSF Quantization Blocks

Spectral LP, suggested by Makhoul [19], is a spectral modelling technique in which the signal spectrum is modelled by an all-pole spectrum. In *selective* (spectral) LP, an all-pole model is applied to a selected portion of the spectrum.

In the case of SBE, the selective LP technique is applied to the HB of the original WB speech, and the spectral envelope of the HB is computed. If, alternatively, a time domain LP analysis is performed on the HB speech, one would need to apply to the WB speech a sharp highpass filter and down-sampling. The filtering operation is costly and is completely eliminated by working in the frequency domain, using the selective LP technique.

To compute the HB spectral envelope, selective LP on the 3-8KHz frequency range is performed on each frame. The selective LPCs are subsequently converted to LSFs and are quantized using an LSF codebook. An LSF vector quantizer (VQ) codebook was designed by the LBG algorithm [22].

2.2.3 Wideband LPC Codebook and Wideband Synthesis Blocks

The problem of WB spectral envelope computation is stated as follows: Given the selective LPCs (or equivalently LSFs) in the frequency range of 3-8KHz, the task is to find WB LPCs in the frequency range 0-8KHz such that an appropriately defined spectral distance between the selective and WB spectral envelopes will be minimal in the HB frequency range of 3-8KHz.

The spectral envelope shape has no importance in the 0-3KHz range since the reconstructed WB speech, generated at the decoder, uses the transmitted NB speech in that frequency range. Hence, the method suggested here for WB spectral envelope computation is based on creating a 0-3KHz spectral envelope by a symmetric folding (mirroring) of the spectral envelope at the frequency range 3-6KHz (in the DFT domain) about the frequency 3KHz. The folding operation is followed by WB LPCs computation using spectral LP. To generate the WB LPC codebook, for each codeword of the given HB LSF codebook, the spectral envelope is reconstructed, and then the symmetric folding operation followed by WB LPCs computation using spectral LP is performed, resulting in a corresponding WB LPC codeword. The generation of the WB LPC codebook is done once, in the design stage. The HB LSF codebook is used for determining the LSF index for a given HB LSF vector. The same index is used to extract the corresponding WB envelope parameters from the WB LPC codebook. The SBE encoder and decoder store the same WB LPC codebook, and use it to generate the WB spectral envelope from a given index of a quantized HB LSF vector.

2.2.4 Gain Estimation and Gain Quantization Blocks

The computation of the HB gain is done to minimize the spectral distance between the spectral envelopes of the original WB speech signal and the reconstructed WB speech signal, in the 3-8KHz frequency range. The spectral difference between these spectral envelopes originates from two main reasons. First, the artificially extended WB excitation is not identical to the original WB excitation. Second, the WB LPCs obtained from the HB quantized LSFs introduce spectral distortion between the two spectral envelopes.

The HB gain factor, denoted by g_{HB}, should minimize the spectral distance between the HB frequency region of the original WB spectral envelope, $|S_{\mathrm{WB}}(\omega)|$ and the HB frequency region of the reconstructed WB speech spectral envelope, $\left|\tilde{S}_{\mathrm{WB}}(\omega)\right|$, multiplied by the HB gain. The error measure for computing the gain factor g is defined by

$$E_{g_{\mathrm{HB}}} \triangleq \frac{1}{\omega_1 - \omega_0} \int_{\omega_0}^{\omega_1} \left(|S_{\mathrm{WB}}(\omega)| - g_{\mathrm{HB}} \left|\tilde{S}_{\mathrm{WB}}(\omega)\right| \right)^2 d\omega \tag{1}$$

The gain factor is found by setting

$$\frac{\partial E_{g_{\mathrm{HB}}}}{\partial g_{\mathrm{HB}}} = 0. \tag{2}$$

By solving (2), the gain factor is equal to

$$g_{\mathrm{HB}} = \frac{\int_{\omega_0}^{\omega_1} |S_{\mathrm{WB}}(\omega)| \left|\tilde{S}_{\mathrm{WB}}(\omega)\right| d\omega}{\int_{\omega_0}^{\omega_1} \left|\tilde{S}_{\mathrm{WB}}(\omega)\right|^2 d\omega} \tag{3}$$

The computed HB gain is quantized for transmission, using a scalar nonuniform quantizer.

2.3 SBE Decoder Structure

The SBE decoder generates the reconstructed WB speech from the received NB speech signal and the embedded side information. The ensuing description of the decoder structure refers to Figure 4. The side information in each speech frame includes the gain index and the LSF codebook index. In the lower branch, the WB excitation signal is generated from the NB speech signal, using the technique used in the SBE encoder (Figure 3). In the middle branch, the WB LPCs are computed by using the LSF codebook index as a pointer to the corresponding WB LPC codebook. The WB artificial excitation together with the gain parameter and the WB LPCs are used to synthesize the WB speech signal. The HB part of the synthesized WB speech signal is filtered by a HPF, and combined with the interpolated NB speech signal, to produce the reconstructed WB speech signal, $\hat{\mathbf{s}}_{\mathrm{WB}}$.

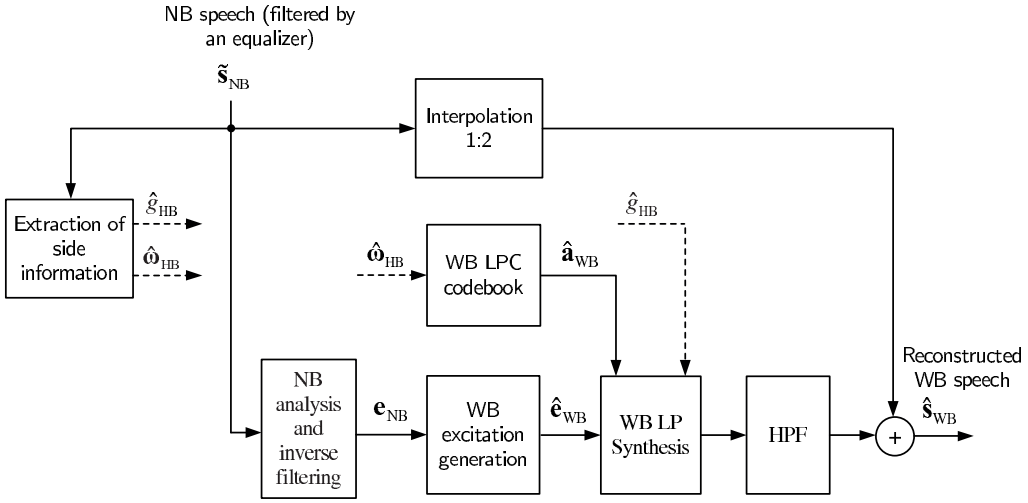

Figure 4: SBE decoder structure.

The input signal to the decoder, denoted by $\tilde{\mathbf{s}}_{\mathrm{NB}}$ in Figure 4, is the output of the channel equalizer. It is desirable that the input to the SBE decoder be as close as possible to the original NB speech signal generated at the input to the telephone channel. Although the NB speech signal which is the output of the channel equalizer is close to the original NB speech, it is not identical to it because of three reasons. First, a residual spectral distortion exists after channel equalization. Second, noise in the transmission channel, which is amplified by channel equalization gets added to the received signal. Third, the existence of embedded data in the NB speech acts like added noise.

3 Perceptual Model-Based Data Embedding

A data embedding (also known as data hiding or digital watermarking) system should satisfy 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 reliably from the watermarked signal, even if it is distorted or attacked. The data embedding *rate* is also of importance in some applications.

In speech and audio coding, a human auditory perception model is used and the irrelevant signal information is identified during signal analysis by incorporating several psychoacoustic principles, such as absolute hearing thresholds, masking thresholds and critical band frequency analysis. Perceptual characteristics of speech and audio coding are incorporated in all modern audio coding standards, such as MPEG audio coders [23]. In data-embedding, the human auditory perception model is used to construct the watermark signal that could be added to the host signal, without affecting the human listener. Auditory perception rules have also been incorporated in SS-watermarking systems [6].

In this section, a method for perceptual model-based data-embedding in speech signals, which combines the SCS technique [3] for data embedding with an auditory masking model, is presented. The proposed encoder performs data embedding in the frequency domain, in separate subbands, utilizing a masking threshold function (MTF). The use of subband masking thresholds (SMTs), derived from the MTF, for the computation of SCS parameters for each subband, is described. Afterwards, the motivation for choosing the discrete Hartley transform (DHT) as the embedding domain is explained. Methods for selecting the subbands for data embedding are also described.

It should be noted that the proposed data-embedding technique, which combines an auditory masking model, is demonstrated here for speech signals but could also be used, with appropriate modifications, for data embedding in audio signals.

We begin the description of the proposed perceptual model-based data-embedding method by presenting the SCS principles in Section 3.1, followed by the description of the subband SCS parameter determination process in Section 3.2. The reasoning for choosing the DHT as the data-embedding domain is given in Section 3.3, and several methods for selecting subbands for data embedding are given in Section 3.4. Finally, the embedded-data decoding process is given in Section 3.6.

3.1 Scalar Costa Scheme Principles

A general model for data communication by data embedding is described in Figure 5. The binary representation of a message m, denoted by a sequence $\mathbf{b}$, is encoded into a coded sequence $\mathbf{d}$ using forward error-correction channel-coding, such as block codes or convolutional codes. The data embedding encoder embeds the coded data $\mathbf{d}$ into the host signal $\mathbf{x}$ producing the transmitted signal $\mathbf{s}$, which is a sum of the host signal $\mathbf{x}$ and the watermark signal $\mathbf{w}$. A deliberate or an unintentional attack, denoted by $\mathbf{v}$, may modify the signal $\mathbf{s}$ into a distorted signal $\mathbf{r}$ and impair data transmission. The data embedding decoder aims to extract 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.

Data Embedding According to SCS [3], the transmitted signal elements are additively composed of the host signal and the watermark signal, i.e.,

$$s_n = x_n + w_n = x_n + \alpha q_n. \tag{4}$$

The watermark signal elements are given by $w_n = \alpha q_n$, where α is a scale factor and q_n is the quantization error of the host signal element quantized according to the data d_n

$$q_n = Q_\Delta \left\{ x_n - \Delta \left(\frac{d_n}{D} + k_n \right) \right\} - \left(x_n - \Delta \left(\frac{d_n}{D} + k_n \right) \right). \tag{5}$$

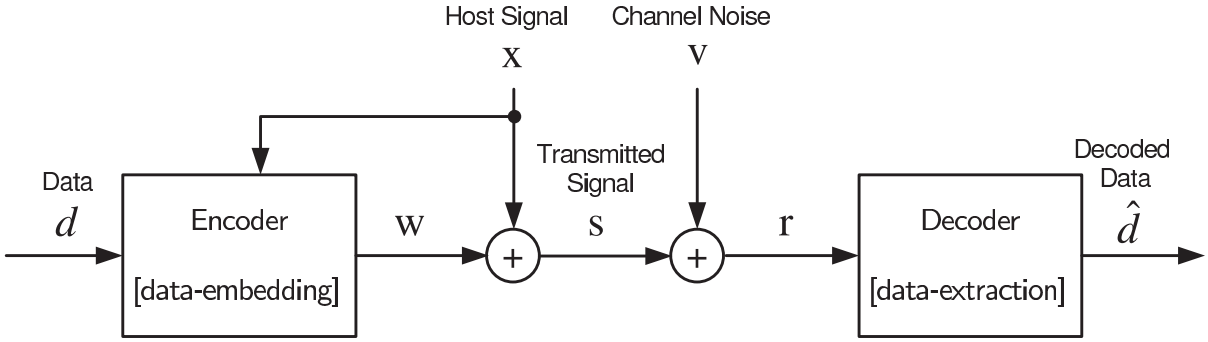

Figure 5: A general model for data communication by data embedding

$Q_\Delta\{\cdot\}$ in (5) denotes scalar uniform quantization with a step-size Δ, and $k_n \in [0,1)$ denote the elements of a cryptographically secure pseudo-random sequence $\mathbf{k}$. For simplicity, it is assumed in the following that the sequence $\mathbf{k}$ is not in use, i.e. $k_n \equiv 0$. The alphabet size is denoted by D. In this paper a binary SCS is utilized, i.e., a SCS with an alphabet size of $D = 2$, and $d_n \in \mathbb{D} = \{0,1\}$, are elements of the data sequence $\mathbf{d}$. The noise elements are given by $v_n = r_n - s_n$, and the watermark-to-noise ratio (WNR) is defined as

$$\mathrm{WNR} = 10\log_{10}\left(\frac{\sigma_w^2}{\sigma_v^2}\right)[\mathrm{dB}], \tag{6}$$

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, and under the assumption of fine quantization, these two parameters are related via

$$\sigma_w^2 = \frac{\alpha^2\Delta^2}{12}. \tag{7}$$

In [3] an analytical expression that approximates the optimum value of α, in the sense of maximizing the capacity of SCS, is given by

$$\alpha_{\mathrm{SCS,approx}} = \sqrt{\frac{\sigma_w^2}{\sigma_w^2 + 2.71\sigma_v^2}}. \tag{8}$$

(7) and (8) lead to $\Delta_{\mathrm{SCS,approx}} = \sqrt{12(\sigma_w^2 + 2.71\sigma_v^2)}$.

Data Extraction In the decoder, data extraction is applied to a signal $\mathbf{y}$, whose elements are computed from the received signal elements r_n by

$$y_n = Q_\Delta\{r_n\} - r_n. \tag{9}$$

Since $|y_n| \le \Delta/2$, y_n is expected to be close to zero if $d_n = 0$ was embedded, and close to $\pm\Delta/2$ if $d_n = 1$. Hence, for proper detection of binary SCS data-embedding, a hard decoding rule should assign

$$\hat{d}_n = \begin{cases} 0 & |y_n| < \Delta/4 \\ 1 & |y_n| \ge \Delta/4 \end{cases}. \tag{10}$$

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 sequence $\hat{\mathbf{b}}$, from the signal $\mathbf{y}$.

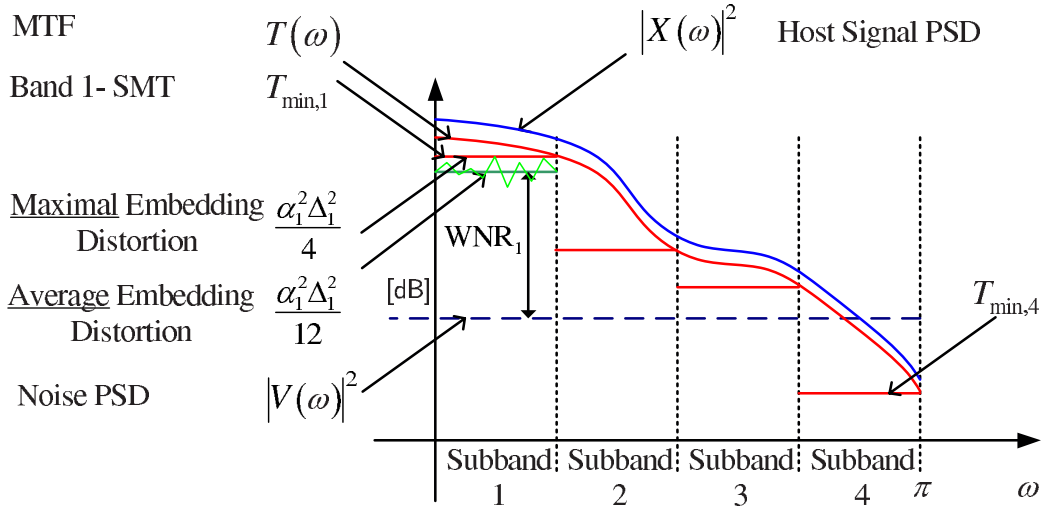

Figure 6: A schematic drawing of a speech signal power spectrum estimate, $|X(\omega)|^2$, divided into 4 subbands; MTF - $T(\omega)$; The SMTs - $T_{min,m}$, are marked by the horizontal solid lines. AWGN source power spectrum estimate $|V(\omega)|^2$ is marked by the dashed line. The WNR in the first subband (WNR$_1$) is also marked.

3.2 Determination of Subband SCS Parameters

The following description is supported by Figure 6. The MTF is computed by the MPEG-1 masking model [23], which is designated for MTF computation for audio signals in general, and for speech signals in particular. The MTF, $\{T(k); \quad 0 \le k \le N/2\}$, with k denoting a discrete frequency index, is calculated for each frame of length N. The positive frequency band is divided into M subbands ($M < N/2$). The subbands could be uniform or non-uniform. The subband masking threshold (SMT) in each subband is set to the minimum of the MTF value in that subband

$$T_{min,m} = \min_{k \in m\text{'th subband}} T(k); \quad m = 1, 2, \ldots, M. \tag{11}$$

The *maximal* embedding distortion (watermark variance) according to (4) and (5) is $\alpha^2\Delta^2/4$, while the *average* embedding distortion is $\alpha^2\Delta^2/12$ (7). Distortion in the m'th subband that is greater than the SMT, $T_{min,m}$ (11), 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

$$10\log_{10}[\alpha_m^2\Delta_m^2/4] = T_{min,m} \quad \text{[dB]}, \tag{12}$$

the subband average embedding-distortion can be expressed in terms of $T_{min,m}$ by

$$\sigma_{w,m}^2 = \frac{\alpha_m^2\Delta_m^2}{12} = \frac{10^{T_{min,m}/10}}{3}. \tag{13}$$

Assuming that a channel-noise model or estimation is given, and denoting the model or estimation of noise variance in the m'th subband by $\sigma_{v,m}^2$, the value of the subband scale factor, α_m, is given by (8)

$$\alpha_m = \sqrt{\frac{\sigma_{w,m}^2}{\sigma_{w,m}^2 + 2.71\sigma_{v,m}^2}}. \tag{14}$$

Formally, the subband quantization-step value is given now, from (13), by

$$\Delta_m^* = \frac{2}{\alpha_m} 10^{T_{min,m}/20}. \tag{15}$$

However, to improve the robustness of the quantization-step detection in the decoder, as well as to reduce the computational complexity of the detection, the applied subband quantization step is selected to be one of a finite pre-defined set of quantization-step values, denoted by

$$\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}. \tag{16}$$

The set of quantization steps is sorted in an ascending order. This set of quantization steps will also be known at the decoder. The quantization-step in the m'th subband is obtained by quantizing the above computed Δ_m^* (15) in the log domain (motivated by the logarithmic sensitivity to sound pressure level of the human listener) yielding

$$\Delta_m = 10^{\frac{D_m}{20}}, \tag{17}$$

where

$$D_m \triangleq c \left\lfloor \frac{T_{min,m} + 20\log_{10}[2/\alpha_m]}{c} \right\rfloor, \tag{18}$$

and the constant c is the quantization step of Δ_m^* in [dB]. Note that for $\mathrm{WNR}_m > 10$[dB], $\alpha_m \cong 1$, simplifying (18), used for the computation of Δ_m^* by (17), to

$$D_m \cong c \left\lfloor \frac{T_{min,m} + 6.02}{c} \right\rfloor. \tag{19}$$

Note that if $\alpha = 1$, SCS is equivalent to dither modulation [7].

3.3 Choice of Data-Embedding Domain

For each type of 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 sound signal as the embedding domain. In other words, the frequency domain coefficients of the host signal are modified according to (4),(5). Several alternative transformations were examined, as follows:

Discrete Fourier Transform The *discrete Fourier transform* (DFT) of the signal frame $\mathbf{x}$, is defined by

$$F_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n e^{\left(-j\frac{2\pi}{N}nk\right)}; \quad k = 0, \ldots, N-1. \tag{20}$$

Discrete Cosine Transform The *discrete Cosine transform* (DCT) of the signal frame $\mathbf{x}$, is defined by

$$C_k \;=\; \beta(k) \sum_{n=0}^{N-1} x_n \cos\left(\frac{(2n+1)k\pi}{2N}\right); \quad k = 0, \ldots, N-1 \tag{21}$$

where

$$\beta(k) = \begin{cases} 1/\sqrt{N}, & k = 0 \\ 2/\sqrt{N}, & 1 \le k \le N-1 \end{cases}$$

Discrete Hartley Transform The *discrete Hartley transform* (DHT) [24] of the signal frame $\mathbf{x}$, is defined by

$$X_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n \mathrm{cas}\left(\frac{2\pi}{N} nk\right); \quad k = 0, \ldots, N-1, \tag{22}$$

where $\mathrm{cas}(x) \triangleq \cos(x) + \sin(x)$. As for the DFT, the transform elements are periodic in k with period N.

The DHT coefficients are used here for data embedding, as this transform is preferred by us over the other two frequency-domain representations: the DFT and the DCT.

The DHT is preferred here over the DFT 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 [25], similar to the those used for the computation of the DFT.

The DFT is commonly used for computing the MTF [23]. Yet, the need for complex arithmetic can be completely eliminated by using the direct relation between the DFT and DHT, given by

$$\mathrm{Re}\{F_k\} = \frac{1}{2}\left[X_{N-k} + X_k\right]; \quad \mathrm{Im}\{F_k\} = \frac{1}{2}\left[X_{N-k} - X_k\right], \tag{23}$$

and

$$|F_k|^2 = \frac{1}{2}\left[X_k^2 + X_{N-k}^2\right], \tag{24}$$

where X_k and F_k denote the DHT and DFT of a signal frame $\mathbf{x}$, respectively. Therefore, in the proposed scheme the DHT is calculated to obtain a representation of the signal for data embedding, followed by the direct computation of the MTF.

Although the DCT is also a real transform, it does not provide the same simplicity in computing the MTF as the DHT. Formally, let Φ_F, Φ_C and Φ_X define the transformation matrices, such that

$$\mathbf{F} = \Phi_F \mathbf{x} \tag{25}$$

$$\mathbf{C} = \Phi_C \mathbf{x} \tag{26}$$

$$\mathbf{X} = \Phi_X \mathbf{x} \tag{27}$$

where $\mathbf{x}$ is a column vector containing the frame elements, and the elements of the transformed vectors $\mathbf{F}$, $\mathbf{C}$ and $\mathbf{X}$ are defined in (20),(21) and (22), respectively. If it is required to transform the MTF, computed by a DFT, to the DCT domain, the MTF $\mathbf{T}$ (a vector whose elements are defined in dB) can be inverse transformed into the vector $\mathbf{t}$ by

$$\mathbf{t} = \Phi_F^{-1} 10^{\mathbf{T}/20}. \tag{28}$$

Then, the MTF in the DCT domain, denoted by $\mathbf{T}_C$, can be computed by

$$\mathbf{T}_C = 10 \log_{10}\left(|\Phi_C \mathbf{t}|^2\right) \quad [\mathrm{dB}] \tag{29}$$

Therefore, computation of $\mathbf{T}_C$ require the computation of the MTF by a DFT, followed by the transformation of the MTF to the DCT domain. This operations could be completely avoided by using the DHT domain for the MTF calculation.

3.4 Selecting Subbands for Data Embedding

We have considered various approaches for selecting the subbands for data embedding. Constraints regarding a *fixed* or *variable embedding-rate* affect the number of subbands in each

frame which are used for data embedding. Further constraints can dictate a *fixed* or *dynamic subband-selection*. Table 1 describes the possible options for *fixed/variable embedding-rate* and *fixed/dynamic subband-selection*.

	Fixed embedding rate	Variable embedding rate
Fixed subband selection	yes	no
Dynamic subband selection	yes	yes

Table 1: Subband selection options

For example, in some applications a *fixed embedding-rate* is required. In that case one can select in advance the subbands (*fixed subband-selection*) that will be used for data embedding, and continue to embed data in these subbands even if the WNR in any of the selected subbands is low. This may result, of course, in a high bit-error-rate (BER). A better option, is to dynamically select a fixed number of subbands, but choose those with the maximal estimated WNR over all subbands. The dynamic approach would obviously result in better performance than a fixed subband selection.

Another option is to have a *variable embedding-rate* with *dynamic subband selection*. In this mode, data is embedded in a specific subband only if the estimated WNR in that subband is greater than a given threshold, that is set according to the allowed BER value. If the actual WNR caused by channel noise, matches the estimated WNR, a target BER value can be ensured. However, as the target BER value is lowered, the attainable data rate is lowered too.

3.5 Composition of Subband Coefficients

The m'th subband coefficients are composed of coefficients from positive and negative frequencies, since the same SMT (11) applies for the corresponding positive and negative frequencies. For example, the m'th subband is composed of the following positive and negative frequency coefficients $[X_{k_{m,\text{start}}}, X_{k_{m,\text{start}}+1}, \ldots, X_{k_{m,\text{end}}}, X_{(N-k_{m,\text{end}})}, X_{(N-k_{m,\text{end}}+1)}, \ldots, X_{(N-k_{m,\text{start}})}]$, where $k_{m,\text{start}}$ and $k_{m,\text{end}}$ are the m'th subband positive frequency boundaries, and $0 < k_{m,\text{start}} < k_{m,\text{end}} < \frac{N}{2}$. 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 (4),(5), with the parameters $\{\alpha_m, \Delta_m\}$. If alternatively, the DFT coefficients are used for data embedding, the embedding can be performed by modifying the real and imaginary parts of the positive frequencies coefficients, and the negative frequencies coefficients are generated by the constraint $F_{N-k} = \bar{F}_k$, since the inverse transformed signal is real. The DHT coefficients are all real and hence not constrained as the DFT coefficients. Therefore, different data can be embedded in the positive and negative frequencies DHT coefficients, providing the same total of N real coefficients that can be used for data embedding. After data embedding, the DHT coefficients are inverse transformed to obtain the transmitted signal.

3.6 Decoding of Embedded Data

There are many types of both deliberate and unintentional *attacks*, which can affect data-embedding systems. A specific unintentional attack, which is caused by transmitting a speech signal with embedded-data over a *telephone channel*, is considered in this paper. When a speech signal with embedded-data is transmitted over the telephone channel, the first step in the decoder is to compensate the spectral distortion introduced by the channel, using an adaptive equalizer, detailed in subsection 3.6.1. Afterwards, frame synchronization is carried out, based on the computed cross-correlation between the stored training signal and the equalizer output signal. The maximum value of the cross-correlation function is searched for, and it's position is used for determining the start position of the first frame. The DHT is then applied to each

frame of the equalized and frame-synchronized signal in order to transform it to the embedding domain.

The next decoding step is the blind detection of embedding-parameters. Blind detection is needed when the decoder does not know the encoding parameters. In the discussed scheme, detection of embedding-parameters include detection of embedded-data presence in each subband, and the detection of the SCS quantization-step. Detection of embedded-data presence in each subband is needed when the encoder chooses dynamically the subbands for data embedding. The subband SCS parameters are also computed dynamically, according to the MTF, and therefore the subband SCS quantization-step needs also to be determined. Since one of a finite set of step values is used (see (16)), determination of the quantization-step is treated as a detection problem, instead of an estimation problem. A combined *maximum likelihood* (ML) detection of embedded-data presence and quantization-step is proposed in Section 3.6.2.

The result of a detection error in the subband embedded-data presence detection or in the quantization-step detection, is a high BER in the subband where the detection error occurred. Therefore, the embedding-parameters detection performance has great influence on the *robustness*. In order to improve the detection performance, the use of a *parameter protection code* (PPC) is suggested in Section 3.6.3.

The final step in the decoder includes extraction of the channel coded data according to hard decoding (10) or soft decoding rule followed by error correction decoding, which results in the decoded embedded data.

3.6.1 Channel equalization

The speech signal transmitted over the telephone line is distorted and noisy, compared to the original speech signal. Trying to operate the decoder on the distorted speech signal would result in a very high BER. As a solution, a channel equalizer is used to compensate the channels' spectral distortion. In data communication literature, there is a variety of algorithms for channel equalization [26][27][28]. In the development stages of this work, several adaptive algorithms were examined for channel equalization, such as the NLMS and RLS algorithms. An equalizer that performs better, in terms of a lower MSE, will usually result in a lower BER in data decoding. Therefore, the RLS algorithm was preferred although it has higher complexity than the NLMS algorithm.

The NLMS and RLS equalization algorithms typically use a pseudo-random white noise training sequence. Since listening to a white noise signal would certainly annoy the listener at the start of a phone conversation, the training stage of the equalization is done in our system in a way that doesn't annoy the listener. This is achieved by replacing the white noise training signal with a musical signal. The musical training signal can be chosen from one of the listeners favorite pieces of music. One demand from the 'musical' equalization is that the training signal occupies the full telephone band, and thus be similar in this aspect to the white noise training signal. Simulation results are reported in Section 4.3.1

Blind equalization algorithms that avoid the need for a training signal are used for equalizing data communication channels, but to the knowledge of the authors there is no blind equalization algorithm that would perform well in our scenario, where data is implicitly embedded in a much stronger analog host signal.

3.6.2 Maximum Likelihood Detection of Embedding Parameters

If *dynamic subband selection* is applied, the decoder has no prior knowledge of either the subband embedded-data presence or the quantization-step. Therefore, the decoder needs to detect these embedding-parameters. The detection stages are:

Step 1: Quantization-Step Determination If data is embedded in a particular subband, the quantization-step used in the embedding is one of a set of quantization step values (sorted in ascending order), $\{\Delta^0, \Delta^1, \ldots, \Delta^{J-1}\}$, as discussed in subsection 3.2. A *test set* of quantization-steps is chosen from the above set, and the test set indices are denoted by $\mathbb{G}$. The minimal and maximal values of the quantization-steps to be tested are denoted by $\Delta^{\min}$ and $\Delta^{\max}$, respectively

Two methods are suggested for the selection of the largest quantization-step to be tested, $\Delta^{\max}$. In the first method, the largest tested quantization-step is set to be the quantization-step obtained by applying (17) with the MTF computed at the decoder. In the second method, $T_{min,m}$ is substituted by $3\sigma^2_{x,m}$ computed at the decoder, and the largest tested quantization-step is computed by applying (17). The latter approach enables a complexity reduction, since there is no need to compute the MTF at the decoder.

The smallest tested quantization-step can be set to $\Delta^{\min} = \Delta^0$. In order to reduce computational complexity, the smallest tested quantization-step can also be set to the smallest quantization-step possible for a given test set size $\{|\mathbb{G}| = G; \quad G > 0\}$. The test set size G is chosen according to an assumed possible range of quantization step values, measured in dB.

Step 2: Computation of the Demodulated DHT Coefficients Using the test set $\mathbb{G}$ of quantization steps, (9) is applied to the received subband DHT coefficients $R_{m,k}$, to obtain $Y^g_{m,k}$. Explicitly, $Y^g_{m,k}$ is computed by

$$Y^g_{m,k} = Q_{\Delta^g}\{R_{m,k}\} - R_{m,k}; \quad g \in \mathbb{G}, \tag{30}$$

where $R_{m,k}$ is the k'th DHT coefficient of the received signal in the m'th subband, and $Y^g_{m,k}$ is computed by (30) from the received DHT coefficient by using each one of the quantization-steps, Δ^g, in the test set $\mathbb{G}$.

Step 3: Computation of Log-Likelihood Ratios In this step, two possible hypotheses are defined, and the log-likelihood ratios (LLRs) are computed from $Y^g_{m,k}$. For notational simplicity, $Y^g_{m,k}$ is replaced by Y, in the next paragraph. The two hypotheses are

- H_0: Y in (30) is computed with the correct quantization-step.
- H_1: Y is computed with the incorrect quantization-step.

The PDFs of the two above hypotheses, $p(Y|H_0)$ and $p(Y|H_1)$, are known at the decoder. Details of computation of the PDFs $p(Y|H_0)$ and $p(Y|H_1)$ are given in [3]. The hypotheses are under the assumption that the embedded-data is present in the subband. Computing Y with the incorrect quantization-step is equivalent to the computation of Y in a subband without embedded-data, since the computation of Y with an incorrect quantization-step will result in uniformly distributed values of Y [3]. Therefore, if embedded-data is absent in a given subband, the demodulated values Y, computed by (30), will have the PDF $p(Y|H_1)$.

The LLR, for each quantization-step of the test set $\mathbb{G}$, is computed by

$$L^g_m \triangleq \log\left(\frac{\prod_{k\in m\text{'th subband}} p(Y^g_{m,k}|H_0)}{\prod_{k\in m\text{'th subband}} p(Y^g_{m,k}|H_1)}\right); \quad g \in \mathbb{G}. \tag{31}$$

The computation of the LLR L^g_m in the above equality is under the assumption that $Y^g_{m,k}$ are statistically independent in the index k. This assumption can be justified in the case of fine quantization. The LLR, L^g_m, is a measure of the validity of the assumption that Δ^g is the quantization step used in the encoder, given that embedded-data is present in that subband.

There are cases when the computation of the LLR will result in a high value, although the tested quantization step Δ^g is not the quantization step used in the encoder, denoted by

Δ^*. One such case happens when the tested quantization step value is large compared to the standard deviation of the subband coefficients distribution. The fine quantization assumption is invalid in this case. To avoid this, one of the previously described methods for the selection of the largest quantization-step to be tested, $\Delta^{\max}$, can be applied. Another case is when the quantization grid of the tested quantization step, Δ^g, and the grid of the quantization step used in the encoder, Δ^*, partly coincide by obeying $2^n\Delta^g = \Delta^*; \{n = 1, 2, \ldots\}$. Since with zero noise the extracted coded data (10) is equal to zero, the Hamming distance between the extracted coded data and a parameter protection code, described in subsection 3.6.3, provides an additional measure of likelihood for the tested quantization step.

Step 4: Embedded-Data Presence Detection The maximal LLR from (31), denoted by $L_m^{g^*}$, is used in the following subband embedded-data presence detection rule

$$\mathbb{I}_m = \begin{cases} 1; & L_m^{g^*} > T \\ 0; & L_m^{g^*} \leq T \end{cases}, \tag{32}$$

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 $\mathbb{I}_m = 0$.

Setting the decision threshold, T, to a value higher than zero will result in a lower false positive detection probability and in a higher false negative detection probability. The setting of $T = 0$ was used in our simulations.

Step 5: Quantization-Step Detection This final step is executed if $\mathbb{I}_m = 1$ in the previous step. The quantization-step in the m'th subband is determined as the quantization-step value that maximizes the LLR, i.e.,

$$\hat{\Delta}_m = \Delta^{g^*}, \tag{33}$$

where

$$g^* = \arg\max_{g \in \mathbb{G}} L_m^g. \tag{34}$$

3.6.3 Parameter Protection Code

The parameter protection code (PPC) can be used to improve the embedded-data presence and quantization-step detection. The PPC is a fixed code, of length N_p, known to the encoder and the decoder, and is denoted by $\{p_n; \quad 0 \leq n \leq N_p - 1\}$. The PPC is appended to the coded data, and embedded in each subband where data is embedded.

For each subband, the decoder computes by hard decoding (10) the decoded PPC, $\hat{p}_n^g$, for each tested quantization step $\{\Delta^g; \quad g \in \mathbb{G}\}$. The decoder computes the Hamming distance, denoted by d_p^g, between the decoded PPC and the original PPC:

$$d_p^g = \sum_{n=0}^{N_p-1} |p_n - \hat{p}_n^g|; \quad g \in \mathbb{G}. \tag{35}$$

As in Section 3.6.2, two possible hypotheses are defined:

- H_0: The decoded PPC is computed with the correct quantization-step.
- H_1: The decoded PPC is computed with the incorrect quantization-step.

The uncoded BER[1], given hypothesis H_0, is denoted by $P_e^{H_0}$, and the uncoded BER, given

[1] *Uncoded BER* is the normalized Hamming distance between the embedded bits, d, and the extracted bits, $\hat{d}$. The *coded BER* is the normalized Hamming distance between the information bits and the decoded information bits.

hypothesis H_1, is denoted by $P_e^{H_1}$. It is assumed that the decoder has prior knowledge on the probability $P_e^{H_0}$, which is dependent on the channel conditions. It is also assumed that $P_e^{H_1} = 1/2$. The probability that the distance between the original and decoded PPC is equal to d_p is given by

$$P(d_p|H_0) = \binom{N_p}{d_p}(P_e^{H_0})^{d_p}(1-P_e^{H_0})^{(N_p-d_p)}, \tag{36}$$

and

$$P(d_p|H_1) = \binom{N_p}{d_p}(P_e^{H_1})^{N_p}. \tag{37}$$

The PPC LLR is defined by

$$P_m^g \triangleq \log\left(\frac{P(d_p|H_0)}{P(d_p|H_1)}\right) = \log\left(\frac{(P_e^{H_0})^{d_p}(1-P_e^{H_0})^{(N_p-d_p)}}{(P_e^{H_1})^{N_p}}\right) \tag{38}$$

Basically, steps 4-5 of the previous section can now be performed, by replacing the LLRs calculated from Y values (31), by the LLRs calculated from the PPC (38). A better option is to combine the two LLRs, as described below.

Combining the LLRs The LLRs calculated from Y values in (31), denoted L_m^g, and the LLRs calculated from the PPC in (38), denoted P_m^g, can be combined for the data-embedding presence and quantization-step detection. There are many ways of combining the above LLRs. A simple combination is to sum the two values,

$$L_{m,\text{combined}}^g = L_m^g + P_m^g, \tag{39}$$

and to use the combined LLR for embedding-parameters detection.

4 Experimental Results

The experimental results reported here are divided into three parts. First, in Section 4.1, we demonstrate the bandwidth extension of telephone-speech, then we detail the telephone channel equalization in Section 4.2, and finally we describe the data-embedding experimental results in Section 4.3.

Subjective listening tests were performed using utterances from the TIMIT database. The subjective tests include a mean opinion score (MOS) evaluation of reconstructed WB speech, a MOS evaluation of NB speech with embedded data, and a preference test between the reconstructed WB speech and the conventional telephone speech. Objective experiments were done using the same database. The results were evaluated by averaging over 625 sentences, having a total duration of more than 34 minutes of speech.

Channel Models Three channel models were used in our simulations: (i) *Telephone channel model* based on the "ITU-T V.56bis" standard [29], which causes amplitude and phase distortion, combined with PCM quantization noise and AWGN. (ii) *PCM channel model* that contains μ-law quantization noise (8 bits/sample), without the telephone channel, and (iii) AWGN channel model with a SNR of 35dB.

4.1 Speech Bandwidth Extension

In our evaluation of the SBE system, we applied an energy-based voice activity detector (VAD) in the SBE encoder to determine in which frames the reconstruction of WB speech should

be performed. In those frames, the SBE encoder computes the HB parameters, and the HB parameters are embedded in the NB speech, as described earlier.

For each input WB signal frame, identified by the VAD as containing speech, the encoder computes and transmits the indices of the HB gain and spectral envelope parameters. The allocation of 12 information bits in a data-embedded subband is divided into 4 bits for the gain index and 8 bits for the LSF index. The NB LP analysis window is of length of 32msec (256 samples at 8KHz sampling rate), but the analysis is updated every 16msec (i.e. with 50% overlap), so that there are two HB updates in each 32msec. A rectangular window is used for extracting frames for data embedding. The DHT coefficients of each non-overlapping frame of 32msec are partitioned into subbands as described in Section 4.3. To support the required SBE side-information, 24 information bits are used in each frame (12 bits in each of the two selected subbands for embedding), resulting in a coded-data rate of 64 bits (two subbands, with 32 bits in each) for each frame. That is, 20 bits are used in each of those two subbands for error correction/protection.

4.1.1 SBE Experiment Results

The proposed SBE system was evaluated by both subjective and objective measures. A subjective MOS test was conducted on 2 sets of 10 sentence long utterances. The first set included WB speech utterances taken from the TIMIT database recordings. The second set comprised reconstructed WB speech utterances generated by passing the first set thorough the complete system (i.e. data embedding, telephone channel, equalization, data extraction and HB reconstruction). 12 non-professional listeners listened to the utterances and rated them on a [1-5] scale: 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. The MOS of the original WB speech was 4.133 and the MOS of the reconstructed WB speech was 3.775. The MOS of the original WB speech utterances is lower than the maximum score of 5 since TIMIT database recordings are intended for the development and evaluation of automatic speech recognition systems, and do not really have excellent quality. The reconstructed WB speech has lower quality than the original WB speech because of two reasons. First, the NB part of the reconstructed speech is noisy, because of the transmission and equalization of the NB speech. Second, the reconstructed HB part is generated from an artificial excitation and the decoded HB parameters.

As in other works [8][17], objective results were evaluated by the log spectral distance (LSD) measure. The objective tool for perceptual evaluation of speech quality (PESQ) [30] in its WB version could perhaps be used for quality evaluation, but an operational WB PESQ software for a 16KHz sampling rate is not at our disposal. The averaged LSD obtained, supported by a side-information rate of 600 bits/sec and measured over the 3.4-7KHz range, was 2.8dB for the simulated telephone channel model. In comparison, in [17], a different structure of the SBE system, which does not use data-embedding, is proposed. In the SBE of [17], the power spectrum is directly vector quantized, in the log domain, requiring a side information rate of 500 bits/sec. The average LSD reported in [17] is 3.6dB, measured over the 3-8KHz range. In the SBE of [8], an LSD of approximately 2.9dB, measured over the 3.4-7KHz range, is supported by a data-rate of 300 bits/sec. However, the result in [8] was obtained with a PCM channel model. With this simplified channel model our suggested system achieved an LSD of 2.6dB at the expense of a higher side information rate. For the AWGN channel model our suggested system obtained an LSD of 2.7dB. The LSD comparison above is under the restriction of not having the same underlying data and applied LSD measure as [8] and [17].

Results obtained for a sample sentence are shown in Figure 7. The original WB speech signal spectrogram is shown in Figure 7(a). The spectrogram of the speech signal filtered by the telephone channel is shown in Figure 7(b) and the reconstructed WB speech signal spectrogram is shown in Figure 7(c). The spectra and spectral envelope of a sample frame of the original and reconstructed WB signals is shown in Figure 8. It can be observed that the NB parts of the spectral envelopes are almost identical, as expected. The difference between these spectral

envelopes is due to imperfect channel equalization. It can also be seen that the HB parts differ more because of the artificial reconstruction process, but this difference was hardly noticed in informal listening.

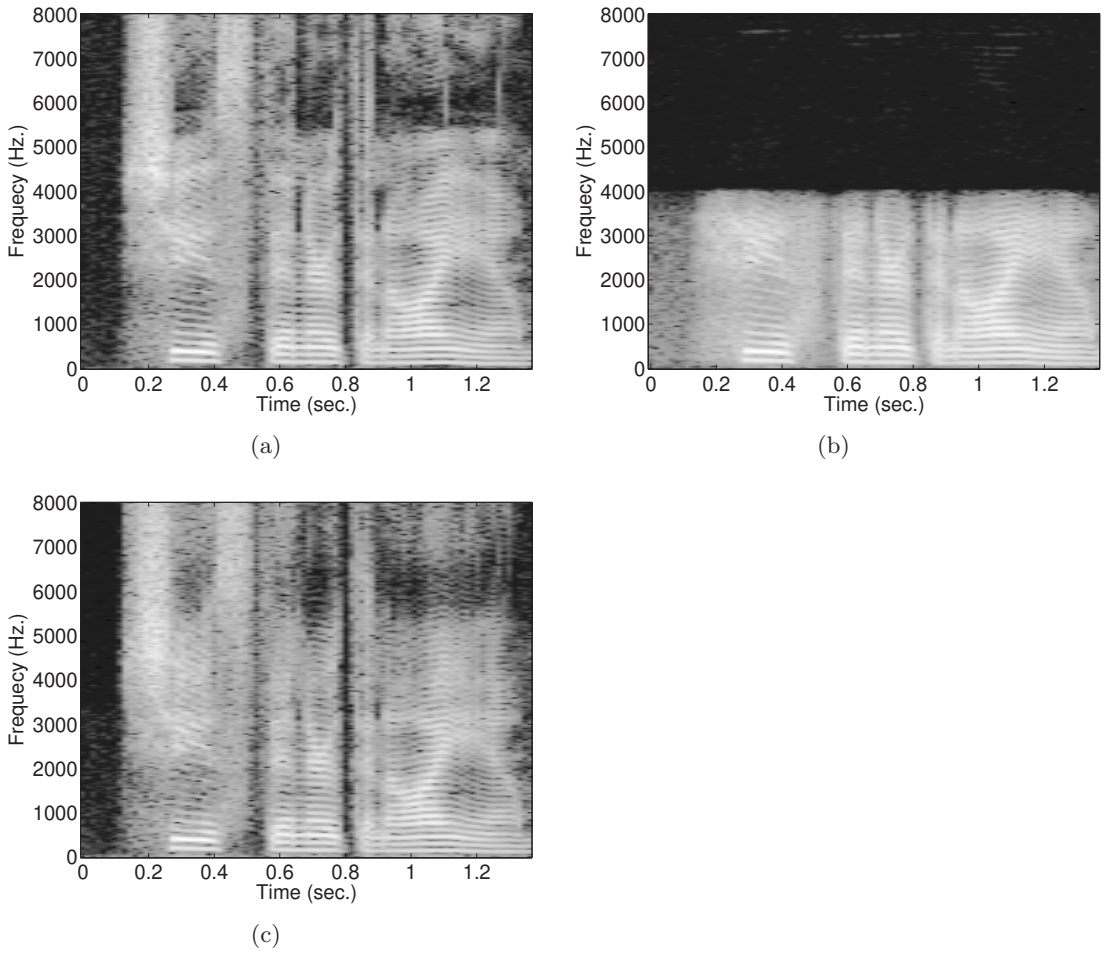

Figure 7: Spectrograms of (a) original WB signal, (b) NB signal, (c) reconstructed WB signal.

Effect of BER on Reconstructed WB Speech Quality In this experiment the data needed for SBE is transmitted by an external side-information channel, and is not embedded in the NB speech. Uniformly distributed random errors were inserted to the side-information bitstream. The channel model is also removed and the SBE encoder and decoder operate in cascade. The LSD as a function of the inserted BER is shown in Figure 9. It can be seen that a BER below 10^{-3} doesn't practically affect the LSD that is achieved by the SBE algorithm. At this BER the LSD is 2.5dB. With a telephone channel model, we obtained a BER of $3.1 \cdot 10^{-4}$ and only a somewhat higher LSD value of 2.8dB, showing that the effect of embedded data noise, channel noise, and remaining spectral distortion after equalization amounts in our system in an increase of 0.3dB only in LSD.

4.2 Telephone Channel Equalization

The RLS algorithm was applied with 256 taps for equalizing the telephone channel. The length of the training sequence is 2^{15} samples, which is approximately 4 seconds long at a sampling rate of 8KHz. Equalization using a musical training signal was also successfully experimented, utilizing part of a classical music piece of Smetana. The averaged LSD obtained with musical

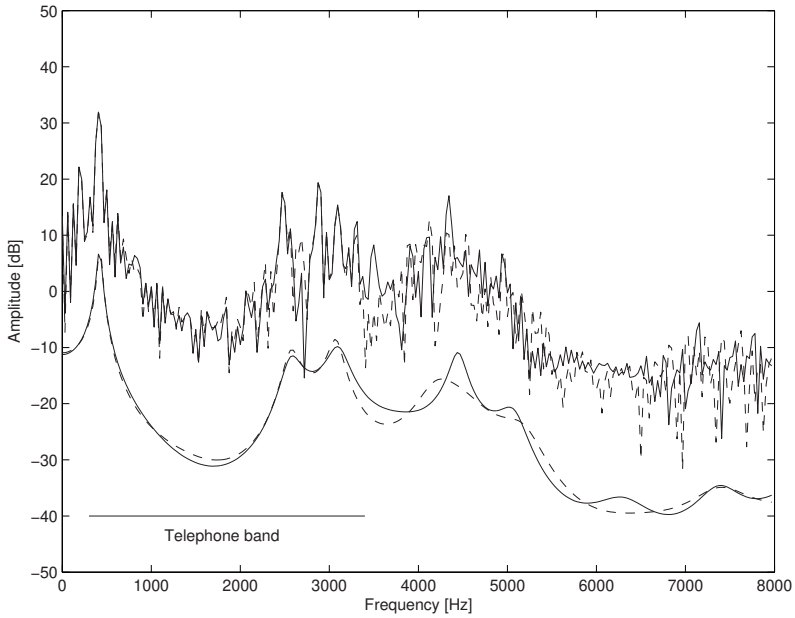

Figure 8: Spectra and spectral envelopes (with a -25 dB offset for display purposes) of original WB speech signal (solid line) and reconstructed WB speech signal (dotted line) produced by SBE decoder.

equalization was 2.8dB, about the same LSD as in the case of a white noise training signal.

4.3 Perceptual Model-Based Data Embedding

As discussed earlier, the computation of the MTF is based on MPEG's psycho-acoustic model [23]. The standard supports several common sampling frequencies of audio signals. Some modifications in the masking model implementation were made in order to suit the case of speech signals sampled at 8KHz.

Since the telephone channel has a large attenuation in the frequency ranges of 0-300Hz and 3400-4000Hz, the full band is partitioned into $M = 8$ non-uniform subbands, as follows: From each frame containing 256 DHT coefficients, the positive and negative frequency coefficients of the first subband, (0-312.5Hz, with the start and end indices of the first subband positive frequency boundaries equal to $k_{0,\text{start}} = 0$ and $k_{0,\text{end}} = 10$, respectively), and of the positive and negative frequency coefficients of the last subband, (3343.75-4000Hz, with the start and end indices of the last subband positive frequency boundaries equal to $k_{7,\text{start}} = 107$ and $k_{7,\text{end}} = 128$, respectively), are not used for data-embedding. The frequency range 343.75-3312.5Hz (with the corresponding start and end indices of the positive frequency boundaries equal to $k_{1,\text{start}} = 11$ and $k_{1,\text{end}} = 106$) is divided into 6 equal width subbands, with each subband containing 32 coefficients from the positive and negative frequencies as described in Section 3.5. From the six subbands, two subbands having the maximal estimated subband WNR were dynamically chosen for data-embedding in each frame, which is detected as containing speech by the VAD. The subband embedded-data is divided into two parts: error corrected coded data and parameter protection code (PPC). A (23,12) Golay block code [28] is used as the error correction code (ECC) for the coded data part, and the PPC part contains a PPC of length $N_p = 9$, $\mathbf{p} = [1, 1, 0, 1, 1, 0, 1, 0, 1]$. Thus, each data embedded subband contains 12 information bits, out of the allocated 32 bits. The average information embedding rate obtained was 600bps. This rate is obtained by multiplying the embedded 24 information bits per frame by the number of frames per second (8000/256) and then by the average VAD rate (0.8).

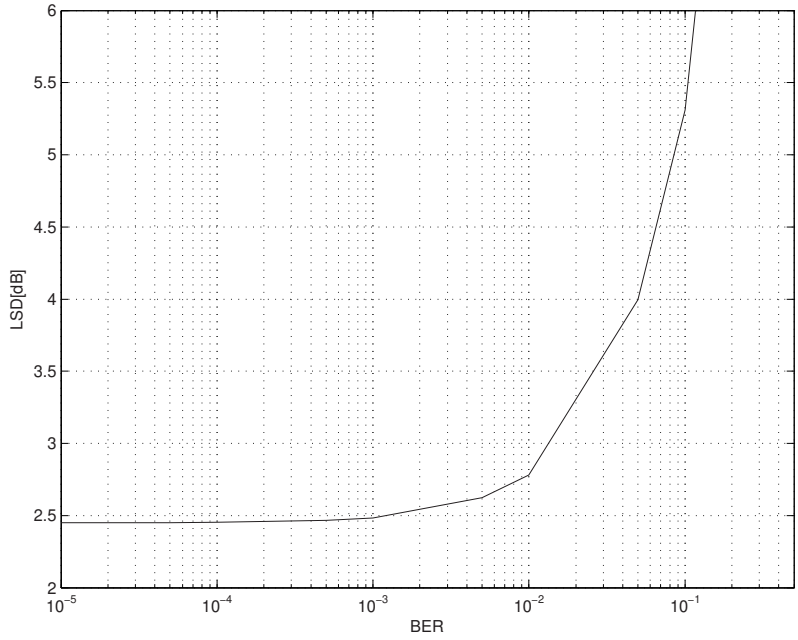

Figure 9: Effect of BER in side information on the SBE LSD.

4.3.1 Data-Embedding Experiments Results

Data Embedding Robustness Robustness of the full system that includes the combined LLRs (39) is described here. Using more than 10^6 information bits, the simulation resulted in the uncoded BER was $9.6 \cdot 10^{-4}$ and the coded BER (following ECC using Golay code) was $3.3 \cdot 10^{-4}$. Detection errors occur when a wrong quantization step is detected in a subband with embedded data, or when a subband without embedded data is detected as containing data. The detection error rate is defined by the ratio of detection errors to the total number of subbands with embedded data. The detection error rate was approximately $4.6 \cdot 10^{-4}$. The utilization of a different ECC is not expected to change significantly the coded BER, since this BER is dominated by the detection error rate.

The embedding scheme of [10] is robust to μ-law quantization noise. In the case of AWGN channel model with a SNR equal to 35dB and an embedding-rate of 216 bits/sec the achievable BER in [10] was 10^{-3}. In the proposed system the embedding-rate is 600 bits/sec and the achievable BER was $3.2 \cdot 10^{-4}$ for the same channel model. For the PCM channel model the achievable BER by our system was $1 \cdot 10^{-4}$.

Data Embedding Transparency Data embedding transparency was evaluated both subjectively and objectively. A subjective MOS test was conducted again on 2 sets of 10 utterances. The first set included NB speech utterances, obtained by a 2:1 decimation of the WB database utterances. The second set comprised the same set of NB speech samples with embedded-data. Both sets were taken before transmission over any channel. 12 non-professional listeners listened to the samples and rated them on a [1-5] scale. The MOS of the NB speech was 3.7 and the MOS of the NB speech with embedded data was 3.625. The small difference between the MOS results demonstrate the transparency of the proposed data embedding scheme. Transparency was evaluated objectively by the PESQ tool for a 8KHz sampling rate. The evaluation results are assumed to be equivalent to a MOS scale of [0-4.5]. Similar to the subjective transparency test, the comparison is between the NB speech and the NB speech with embedded-data. The PESQ score result, averaged over 625 sentences, was approximately 3.9.

The authors of [10] conducted a subjective test, in which they asked participants to compare

the NB speech and the NB speech with embedded data by a four grade scale: 1-the two signals are quite different; 2-the two signals are similar, but the difference is easy to see; 3-the two signals sound very similar, little difference exists; 4-the two signals sound identical. The subjective test result was 3.07.

A "nearly imperceptible watermark" was reported in [8], while no numerical objective or subjective measures were given.

4.3.2 Subjective Comparison of Reconstructed WB Speech and Telephone Speech

In order to examine the complete scheme of bandwidth extension of telephone speech aided by data embedding, an A-B preference test was conducted by the same 12 non-professional listeners as in the previous MOS tests. The participants were asked to compare the quality of A-B utterance pairs, and to rate if the quality of one is *much better*, *better*, or *the same*, compared to the other utterance. Conventional telephone speech utterance without embedded data was compared to the reconstructed WB signal, created by the complete scheme. The results are summarized in Table 2. Note that the proposed system achieved 92.5% preference, at different degrees, over the conventional telephone speech.

Preference	Same	Reconstructed WB speech (set A)		Telephone speech (set B)	
		A is better	A is much better	B is better	B is much better
%	3.33	67.5	25	3.33	0.83

Table 2: A-B preference test for the reconstructed WB speech and the conventional telephone speech.

5 Conclusion

We have presented a system for bandwidth extension of telephone-speech aided by data embedding. The proposed system uses the transmitted NB speech signal as a carrier of the side information needed to carry out the bandwidth extension, thus eliminating the need for an additional channel. We have also proposed a novel data embedding scheme, in which the scalar Costa scheme is combined with an auditory masking model allowing high rate transparent embedding at a low bit error rate. The embedded data payload can also be used for purposes other than SBE. For example, text and graphics can be transmitted as embedded-data during an ongoing conversation. Subjective tests showed that the WB speech output of the suggested SBE system was preferred (at different degrees) over conventional telephone speech in 92.5% of the test utterances. In another listening test the MOS of the NB speech was 3.7 and the MOS of the NB speech with embedded data was 3.625. The small difference between the MOS results demonstrate the transparency of the proposed data embedding scheme. In simulations, the embedded data rate was 600 information bits/second with a bit-error-rate of approximately $3 \cdot 10^{-4}$. The averaged LSD obtained, measured over the 3.4-7KHz range, was 2.8dB. Further details regarding the suggested SBE system supported by data-embedding can be found in [31].

Future work may be directed to various components of the proposed system:

Embedding-rate improvements: (i) It was shown in [3], that binary SCS capacity is limited for high WNRs due to the binary alphabet of embedded-data letters. Throughout this work binary SCS was utilized. Since the experimental average subband WNR is high, approximately 18dB, the rate can be increased by applying D-ary SCS, with $D > 2$. (ii) *Lattice Costa scheme* [32], which employs lattice quantization instead of scalar quantization, can also be used for embedding-rate improvement. (iii) In the suggested application, only two subbands are used for data embedding in each frame. The encoder chooses these subbands as the ones with the highest estimated WNR for each frame. The embedding rate could be increased by dynamically

choosing also the number of subbands for data-embedding, from the set of subbands into which the transformed signal frames are divided.

Blind channel equalization: The examined algorithms for channel equalization make use of a training sequence for the adaption stage. If blind channel equalization could be used, this stage could be avoided. Developing a blind channel-equalization algorithm for data embedding systems appears to be a challenge.

Acknowledgments

The authors wish to thank the students of the Signal and Image Processing Lab (SIPL) who volunteered to participate in the listening tests.

References

[1] S. Voran, "Listener ratings of speech passbands," in *Proc. of the 1997 IEEE Workshop on Speech Coding for Telecommunications*, Pocono Manor, PA, September 1997, IEEE, pp. 81–82.

[2] P. Jax and P. Vary, "Bandwidth extension of speech signals: a catalyst for the introduction of wideband speech coding?," *IEEE Communications Magazine*, vol. 44, pp. 106–111, May 2006.

[3] J. J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1003–1019, April 2003.

[4] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[5] Q. Cheng and J. Sorensen, "Spread spectrum signaling for speech watermarking," in *Proc. of ICASSP*. IEEE, 2001, pp. 1337–1340.

[6] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337–355, 1998.

[7] 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*, vol. 44, no. 4, pp. 1423–1443, May 2001.

[8] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *9th European Conf. on Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, September 2005, pp. 1497–1500.

[9] A. Sagi and D. Malah, "Data embedding in speech signals using perceptual masking," in *European Signal Processing Conference*, Vienna, Austria, September 2004, pp. 1657–1660.

[10] S. Chen and H. Leung, "Concurrent data transmission through analog speech channel using data hiding," *IEEE Signal Processing Letters*, vol. 12, no. 8, pp. 581–584, August 2005.

[11] E. Larsen and R. M. Aarts, *Audio Bandwidth Extension*, Wiley, 2004.

[12] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 266–274, 2001.

[13] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[14] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Speech Processing*, vol. 83, pp. 1707–1719, 2003.

[15] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model," in *Proc. of ICASSP*, Istanbul, 2000, IEEE, vol. 4, pp. 1153–1156.

[16] A. McCree, T. Unno, A. Anandakumar, A. Bernard, and E. Paksoy, "An embedded adaptive multi-rate wideband speech coder," in *Proc. of ICASSP*, Salt-Lake City, UT, May 2001, IEEE, vol. 2, pp. 761–764.

[17] J. M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," *IEEE Workshop on Speech Coding*, pp. 130–132, September 2000.

[18] J. R. Epps and W. H. Holmes, "A new very low bit rate wideband speech coder with a sinusoidal highband model," *IEEE International Symposium on Circuits and Systems*, pp. II349–II352, September 2001.

[19] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 3, pp. 283–296, June 1975.

[20] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communications Systems*, Wiley, 1994.

[21] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. of ICASSP*, Washington, DC, 1979, pp. 428–431.

[22] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[23] 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," Tech. Rep. ISO/IEC 11172-3, ISO, 1992.

[24] R. N. Bracewell, "Discrete Hartley transform," *J. Opt. Soc. Am.*, vol. 73, no. 12, pp. 1832–1835, 1983.

[25] H. Sorensen, D. Jones, C. Burrus, and M. Heideman, "On computing the discrete Hartley transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 5, pp. 1231–1238, October 1985.

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

[27] S. Haykin, *Communication Systems*, John Wiley & Sons, Inc., 4th edition, 2001.

[28] B. Sklar, *Digital Communications, Fundemantals and Appplications*, Prentice-Hall, 1988.

[29] ITU-T, "Netwrok transmission model for evaluating modem performance over 2-wire voice grade connections," Tech. Rep. V.56 bis, ITU, August 1995.

[30] 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," Tech. Rep. P.862, ITU, February 2001.

[31] A. Sagi, "Data embedding in speech signals," M.S. thesis, Technion-Israel Inst. of Technology, Haifa, Israel, May 2004.

[32] R. F. H. Fischer, R. Tzschoppe, and R. Bauml, "Lattice costa schemes using subspace projection for digital watermarking," in *5th International ITG Conference on Source and Channel Coding (SCC)*, Erlangen, Germany, January 2004, pp. 127–134.