DELAYLESS SOFT-DECISION DECODING OF HIGH-QUALITY AUDIO TRANSMITTED OVER AWGN CHANNELS

Florian Pflug and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
{f.pflug.t.fingscheidt}@tu-bs.de

ABSTRACT
Short-range wireless audio transmission with high quality on the one hand often encounters error-prone channels, while on the other hand decoding delay plays a critical role in the application. A lot of prior art in audio error concealment is either only intuitively motivated or adds too much delay to the transmission. In this paper we propose a framework for transmitted audio error concealment without any algorithmic delay. As a novelty, it strongly follows a Bayesian approach for quantized but uncompressed audio, which in principle be applied to any type of wireless channel yielding some kind of reliability information. Two methods to compute audio prediction coefficients are presented, one based on the autocorrelation method, the other one on the normalized least-mean-square (NLMS) algorithm. Simulation results for additive white Gaussian noise (AWGN) channels show the significant effect of the proposed approaches.

Index Terms— delayless soft-decision audio decoding, linear prediction, normalized least-mean-square algorithm, autocorrelation, Levinson-Durbin algorithm

1. INTRODUCTION
In digital communications robust source decoding algorithms are necessary wherever transmission errors occur. This is due to the fact that even single residual bit errors after a possible channel decoding scheme can lead to heavy distortions when they affect critical parts of a coded signal, e.g., the most significant bits of a pulse-code modulated (PCM) audio sample. Therefore, efficient error concealment strategies are essential.

Recently, robust decoding algorithms exploiting residual redundancy in source coded signals have found increasing attention (see, e.g., [1–4]). However, the investigated signals are either often narrowband speech signals with rather low sampling rates (8 kHz) and coarse quantization or audio signals largely delayed due to the use of efficient source coding algorithms (e.g., MPEG-1 Audio Layer 2 or MPEG-4 Part 3 AAC). In addition, with upcoming digital audio transmission systems such as WirelessHD [5], Wireless Home Digital Interface (WHDI) [6] or digital wireless head- and microphones the demand for digital transmission of high-quality audio signals with very low delay rises. As a result, these systems transmit the audio signal without computationally complex source coding or even in uncompressed form. Furthermore, the signals are mostly of CD quality and are therefore sampled at 44.1 or 48 kHz and quantized with 16 bits per sample. For professional grade and audiophile systems even sampling rates of 96 and 192 kHz and 24 bit quantization are common. With known approaches such as [2–4, 7] an enormous increase in decoding complexity would result.

In order to achieve a sufficient perceptual quality for distorted audio signals, intensive research has been conducted in the area of error concealment, in particular denoising, declicking and decrackling (see, e.g., [8–12]). However, from the point of view of source decoding, these algorithms work on signals generated by hard-decision. Therefore, they do not benefit from a bit-wise reliability information obtained during demodulation and/or channel decoding. In contrast, our novel approach exploits such bit-wise reliability information together with the residual redundancy of an audio signal to compute an estimate of each received sample without any algorithmic delay. It is based on an approach to robust speech decoding introduced in [13] and is adapted to be suitable for soft-decision decoding of general audio signals with fine quantization and high sample rates in this work. Moreover, we do not employ any channel coding scheme, instead we purely rely on pre-trained a priori knowledge and the inherent redundancy of audio signals. For reasons of simplicity, our investigations focus on short-range line-of-sight scenarios. Therefore, we employ an additive white Gaussian noise (AWGN) channel but our approach can easily be extended to more complex channel models.

The organization of the paper is as follows: Section 2 describes the basic steps of soft-decision audio decoding. In Section 3 our new approach to compute prediction probabilities is presented in detail. Simulation results are discussed in Section 4. Finally, Section 5 draws conclusions with regard to the achieved results.

2. SOFT-DECISION AUDIO DECODING
The high-level framework of our soft-decision audio decoding scheme basically follows the approach presented in [3]. The complete simulation chain as depicted in Fig. 1 starts at sample index \( n \) with audio samples \( s_n \) in the range \( [-2^{M-1}, -2^{M-1} + \Delta, \ldots, 2^{M-1} - \Delta] \) quantized in \( M \) bit resolution, with \( \Delta = 1 \) denoting the quantization step size. These samples are mapped to bit combinations \( x_n = (x_n(0), x_n(1), \ldots, x_n(m), \ldots, x_n(M-1)) \) with bipolar bits \( x_n(m) \in \{-1, 1\} \). After the transmission over an AWGN channel the soft demodulator yields a log-likelihood ratio (LLR)

\[
L(\hat{x}_n(m)) = \ln \frac{P(\hat{x}_n(m)|x_n(m) = +1)}{P(\hat{x}_n(m)|x_n(m) = -1)}
\]

which delivers receiver-side bit error probabilities \( p_{e,n}(m) = 1/(1 + \exp[L(\hat{x}_n(m))]) \) along with each hard-decided bit \( \hat{x}_n(m) = \text{sign}(L(\hat{x}_n(m))) \). The soft-decision audio decoder uses this reliability information together with a priori knowledge for an estimate \( \hat{s}_n \) of the received audio sample. A brief overview of the algorithmic steps is given in the following.
2.1. Transition Probabilities

The channel reliability information \( p_{e,n}(m) \) obtained from the soft demodulator is used for the computation of transition probabilities \( P(\hat{s}_n|x^{(i)}) \). This describes the probability for the known received hard-decided bit combination \( \hat{s}_n \) given a possibly transmitted bit combination \( x^{(i)}, i \in \{0, 1, \ldots, 2^M-1\} \). Due to the memoryless channel it can be obtained by

\[
P(\hat{s}_n|x^{(i)}) = \prod_{m=0}^{M-1} P(\hat{x}_n(m)|x^{(i)}(m)),
\]

with

\[
P(\hat{x}_n(m)|x^{(i)}(m)) = \begin{cases} p_{e,n}(m) & \text{if } \hat{x}_n(m) \neq x^{(i)}(m), \\ 1 - p_{e,n}(m) & \text{else}. \end{cases}
\]

2.2. A Posteriori and A Priori Probabilities

Transition and a priori probabilities are combined to a posteriori probabilities which describe the probabilities of all possibly transmitted bit combinations \( x^{(i)} \) given the received bit combination \( \hat{s}_n \) and its complete history \( \hat{s}_{n-1} = (\hat{s}_{n-1}, \hat{s}_{n-2}, \ldots) \). Thereby, we do not introduce any algorithmic delay because we only rely on present and past bit combinations. Due to the memoryless channel the a posteriori probability reads as

\[
P(x^{(i)}|\hat{s}_n, \hat{s}_{n-1}^{-1}) = \frac{1}{C} \cdot P(\hat{s}_n|x^{(i)}) \cdot P(x^{(i)}|\hat{s}_{n-1}^{-1}),
\]

with \( P(x^{(i)}|\hat{s}_{n-1}^{-1}) \) denoting prediction probabilities and \( C \) being a constant, such that \( \sum_{i=0}^{2^M-1} P(x^{(i)}|\hat{s}_n, \hat{s}_{n-1}^{-1}) = 1 \).

Prediction probabilities \( P(x^{(i)}|\hat{s}_{n-1}^{-1}) \) describe a priori knowledge about a transmitted bit combination \( x^{(i)} \) on the receiver side before the corresponding \( \hat{s}_n \) has been received. As a result, if no a priori knowledge (NAK) is available, the probabilities of all possibly transmitted bit combinations are uniformly distributed and the prediction probability reduces to

\[
P(x^{(i)}|\hat{s}_{n-1}^{-1}) = 2^{-M}.
\]

With the availability of \( 0^\text{th} \) order a priori knowledge the prediction probabilities can be written as

\[
P(x^{(i)}|\hat{s}_{n-1}^{-1}) = P(x_n = x^{(i)}).
\]

This can be gathered in advance from representative training audio data, for example by counting the occurrences of each sample value in the training material. In this case, for audio with a resolution of \( M = 16 \) bits per sample, \( 2^M = 2^{16} \) words have to be stored in the decoder. This method is referred to as AK0 in the following.

Instead of a one-time measurement of \( 2^M \) a priori probabilities in advance, prediction probabilities for soft-decision audio decoding can also be determined on the receiver side by means of linear prediction. As a result, not only the reliability information of the current bit combination \( \hat{s}_n \) is exploited but also the correlation of previously estimated samples. This new approach is detailed in Section 3.

2.3. Sample Estimation

For the final estimation \( \hat{s}_n \) of the transmitted sample \( s_n \) the well-known minimum mean-square-error (MMSE) estimator is utilized. It uses the mean-square error (MSE) \( E[(s_n - \hat{s}_n)^2] \) as an optimization criterion, with \( E[\cdot] \) denoting the expectation value. The MMSE estimation rule can be written as

\[
\hat{s}_n = \sum_{i=0}^{2^M-1} s^{(i)} \cdot P(x^{(i)}|\hat{s}_n, \hat{s}_{n-1}^{-1}),
\]

with \( s^{(i)} \in \{-2^M-1, -2^{M-1}+\Delta, \ldots, 2^{M-1}-\Delta\}, \Delta = 1 \), being all possibly transmitted audio samples in \( M \) bit resolution, represented by bit combination \( x^{(i)} \).

3. NEW APPROACH TO PREDICTION PROBABILITIES

In order to successfully exploit the inherent redundancy of an audio signal, it is necessary to include preceding samples in the computation of prediction probabilities. This can be done by a linear combination of \( N_p \) preceding audio samples \( s_{n-1-N_p} = (s_{n-1}, s_{n-2}, \ldots, s_{n-N_p})^T \) [13], with \( N_p \) being the prediction order and \( (\cdot)^T \) denoting a transposed vector. A linear prediction of \( s_n \) is given by

\[
s_{p,n} = a_n \cdot s_{n-N_p}^n, \tag{8}
\]

with \( a_n = (a_n(1), \ldots, a_n(N_p))^T \) being \( N_p \) predictor coefficients. The error signal of the prediction is denoted by \( e_n = s_n - s_{p,n} \). In a training process, the statistics of these errors is measured, leading to a probability density function (PDF) \( p_E(e) \). Two approaches will be detailed in Sec. 3.1 and 3.2, respectively, where \( a_n \) is updated for each sample instant \( n \) in order to minimize the MSE \( E[e_n^2] \).

In order to implement the predictor at the receiver side as shown in Fig. 2, it is necessary to feed it with previously estimated samples. The prediction value \( \hat{s}_{p,n} = s_{n-N_p}^T \cdot s_{n-N_p}^{-1} \) is by definition of \( a_n \) an MMSE estimate of \( \hat{s}_n \) given \( \hat{s}_{n-1-N_p} \). In addition, each \( s^{(i)} \) and \( s_{n-N_p}^{-1} \) can be bijectively mapped to exactly one \( x^{(i)} \) and \( \hat{s}_{n-N_p}^{-1} \), respectively. Since we also have a direct mapping of \( s_{n-N_p}^{-1} \) to \( \hat{s}_{n-N_p}^{-1} \), the prediction probability can finally be written as

\[
P(x^{(i)}|\hat{s}_{n-N_p}^{-1}) = p_E(e^{(i)} = s^{(i)} - \hat{s}_{p,n}), \tag{9}
\]

This means that the prediction probability is available for each \( s^{(i)} \) by addressing the pre-trained PDF \( p_E(e) \) at \( e = s^{(i)} - \hat{s}_{p,n} \). This is illustrated with the box on the left-hand side of Fig. 2.

\[
\text{Fig. 1. Block diagram of the simulation setup}
\]

\[
\text{Fig. 2. Block diagram for the computation of prediction probabilities}
\]

\[
\text{P(x^{(i)}|\hat{s}_{n-N_p}^{-1})}. \text{Dashed lines indicate paths only necessary for the}
\]

\[
\text{normalized least-mean-square (NLMS) approach.}
\]
3.1. Autocorrelation Approach

A well-known solution for computing predictor coefficients is the autocorrelation (ACOR) method [14]. It starts with windowing of $L$ previously estimated audio samples $\hat{s}_{n-1}$, according to $W \cdot \hat{s}_{n-1}$, with $W = 0.54 \cdot I - 0.46 \cdot \text{diag} \{ \cos(\frac{\pi}{T}), \ldots, \cos(\frac{L \pi}{T}) \}$ being a left-half Hamming window, $I$ being the identity matrix and $\text{diag} \{ \cdot \}$ denoting a diagonal matrix built using the elements of its argument vector. The windowed samples serve to compute the autocorrelation (ACOR) method [14]. It starts with windowing of previously estimated audio samples $\hat{s}_{n-1}$, and for its complexity, the algorithm is linear in the prediction order. Furthermore, the coefficient update is normalized with respect to the squared Euclidean norm $||\cdot||^2$ of $N_p$ previously estimated samples. As a result, the predictor coefficient update is only loosely dependent on the energy of $\hat{s}_{n-2}$. Therefore, this algorithm is well-suited for high-dynamic audio signals. At sample instant $n = 0$ the variables are initialized using $a_{-1} = (1/N_p, \ldots, 1/N_p)^T$ and $\hat{s}_{-1} = (0, \ldots, 0)^T$ [16].

3.2. Normalized Least-Mean-Square Approach

The normalized least-mean-square (NLMS) algorithm is widely used for adaptive filtering [15] and lossless audio source coding [16, 17]. It adaptively estimates filter coefficients according to

$$a_n = a_{n-1} + \frac{\hat{e}_{n-1}}{1 + \lambda \cdot ||\hat{s}_{n-1}||^2} \cdot \hat{s}_{n-1}^T$$  \hspace{1cm} (11)

such that the mean-square error between the filter output $\hat{s}_{p,n}$ and the desired response $s_n$ is minimized, with $\lambda$ being a tuning parameter controlling the convergence rate [16]. The complexity of the algorithm is linear in the prediction order. Furthermore, the coefficient update is normalized with respect to the squared Euclidean norm $||\cdot||^2$ of $N_p$ previously estimated samples. As a result, the predictor coefficient update is only loosely dependent on the energy of $\hat{s}_{n-1}$. Therefore, this algorithm is well-suited for high-dynamic audio signals. At sample instant $n = 0$ the variables are initialized using $a_{-1} = (1/N_p, \ldots, 1/N_p)^T$ and $\hat{s}_{-1} = (0, \ldots, 0)^T$ [16].

4. SIMULATIONS

4.1. Experimental Setup

For our simulations 13 monaural audio signals (excerpts from classical pieces and a motion picture sound track with music and effects) with different musical instruments (strings, organ, brass instruments, percussions, piano, synthesizer), exhibiting a total length of 96 s are transmitted over an AWGN channel according to Fig. 1 with $E_b/N_0$ ratios ranging from 0–12 dB. The $a$ priori probabilities $P(x^{(b)})$ for AK0 decoding and the prediction error PDF $p_E(x)$ for the ACOR and NLMS approaches have been gathered from 15 audio signals comprising classical compositions, electronic music and a motion picture soundtrack including music, speech and effects with a total length of 81 min. This data is exclusively used for training. The global signal-to-noise ratio (SNR) is

$$\text{SNR}_{\text{global}} = 10 \log_{10} \frac{E\{s^2\}}{E\{(s - \hat{s})^2\}}$$  \hspace{1cm} (12)

and the segmental SNR

$$\text{SNR}_{\text{seg}} = \frac{10}{K} \sum_{k=0}^{K-1} \log_{10} \frac{\sum_{n=0}^{N-1} (\hat{s}_{n+k,N} - s_{n+k,N})^2}{\sum_{n=0}^{N-1} (\hat{s}_{n+k,N} - s_{n+k,N})^2}$$  \hspace{1cm} (13)

serve as a measure of audio quality, with $K$ being the number of frames and $N = 480 \pm 10$ ms denoting the frame length. Whereas the global SNR reflects seldom but significant artifacts such as loud clicks, the segmental SNR is an indicator for frequent but low-amplitude artifacts like low-volume noise or crackling (SNR global and SNR seg have been proven to be particularly useful in the context of transmission errors and error concealment schemes [3]). Note that all audio signals are sampled with 48kHz, are 16 bit linear pulse-code modulated (PCM) and have been normalized to –26dBFS (dB.
relative to full-scale). Only for the SNR measurements denoted by Eqs. (12) and (13) the reference signal \( s \) is available in 24 bit resolution. Assuming a perfectly Laplacian-distributed signal at \(-26\,\text{d BFS}\) and fine uniform quantization \( E_\{\{s^2\}\} = \Delta^2/12 \), a maximum global SNR of approximately 77 dB can be achieved.

We conducted a large number of experiments in order to identify well-suited window sizes \( L \), prediction orders \( N_p \), and convergence parameters \( \lambda \) for the ACOR and NLMS approaches, respectively. It turned out that \( \lambda = 20 \) for the NLMS and \( N_p = 10 \) for the ACOR approach are reasonable choices. In the current work, we investigate the parameters \( N_p \in \{10, 240\} \) for the NLMS algorithm and \( L \in \{240, 960\} \) with \( N_p = 10 \) for the ACOR approach.

### 4.2. Discussion

Figs. 3 and 4 present the simulation results for the soft-decision ACOR and NLMS approaches, soft-decision decoding with 0th order \textit{a priori} knowledge (AK0), without \textit{a priori} knowledge (NAK) and hard-decision decoding (HD). The channel \( E_0/N_0 \) ratios are given on the abscissae, the resulting SNRs after decoding on the ordinates.

As can be seen, the novel ACOR and NLMS approaches are consistently superior to all other investigated audio decoding algorithms, reaching gains of up to 8 dB \( E_0/N_0 \) and 40 dB in both SNR\textsubscript{global} and SNR\textsubscript{seg} compared to HD decoding. At very high \( E_0/N_0 \) ratios all decoding algorithms line up for an SNR\textsubscript{global} of 75.1 dB and an SNR\textsubscript{seg} of 72.1 dB, respectively.

Only for the lowest \( E_0/N_0 \) ratio of 0 dB, the AK0 algorithm reaches an SNR\textsubscript{global} performance similar to that of the ACOR and NLMS approaches. This is due to the fact that it is more reasonable for very bad channels to use a correct amplitude distribution \( P(\{x\}) \) than exploiting correlation to unconfidently estimated samples. However, as depicted in Fig. 4, minor artifacts are still better concealed by the ACOR and NLMS approaches. Furthermore, plain HD decoding outperforms NAK decoding in SNR\textsubscript{seg} for \( E_0/N_0 \) ratios greater than 5 dB. This is because high-energy clicks become less likely for rising \( E_0/N_0 \) ratios but the NAK algorithm introduces strong background noise due to uniformly distributed \textit{a priori} probabilities. Please note that informal listening tests confirm our findings from the instrumental SNR measures.

Finally, it should be noted that the light-weight NLMS algorithm with \( N_p = 10 \) consistently delivers remarkably good results. It reaches the SNR\textsubscript{global} performance of the much more computationally demanding ACOR algorithm for \( E_0/N_0 \) ratios above 6 dB. Furthermore, Fig. 4 reveals that it delivers the same SNR\textsubscript{seg} performance as the ACOR algorithm with \( L = 240 \), even outperforming it for low \( E_0/N_0 \) ratios up to 7 dB. This clearly demonstrates that delayless and efficient soft-decision audio decoding is feasible with the low-complexity NLMS approach.

### 5. CONCLUSIONS

In this contribution we have presented a Bayesian framework for efficient error concealment of audio transmitted over wireless channels. Considering channel reliability information, two methods to compute short-term prediction coefficients – the autocorrelation method and an NLMS approach – have been investigated. The predicted audio sample is used to determine a prior distribution of the audio sample required for the Bayesian framework. For an AWGN channel both methods have shown significant gains of up to 8 dB \( E_0/N_0 \) without any channel coding or any additional algorithmic delay. Our approach can be used in a wide field of audio applications such as audio streaming in WirelessHD, Wireless Home Digital Interfaces (WHDI) or in wireless microphones or headphones.

### 6. REFERENCES


