SPEECH WATERMARKING THROUGH PARAMETRIC MODELING

A. Gurijala‡, J.R. Deller, Jr.‡, M.S. Seadle∗

Michigan State University
‡Dept. Elec. & Computer Engr. / 2120 EB
∗MSU Libraries / 308 LIB
East Lansing, MI 48824 USA

John H.L. Hansen
University of Colorado
Center for Spoken Language Research
Boulder, CO 80309 USA

ABSTRACT

A general formulation for speech watermarking through parametric modeling is suggested, then the paper focuses on a watermarking technique based on linear-predictive (LP) modeling of speech. In the particular strategy employed here, information is embedded by modifying the autocorrelation values of the original speech. The amount of information that can be embedded is subject to fidelity constraints. The modified LP coefficients derived from the new set of autocorrelation values are used for reconstructing the watermarked speech. The perceptual quality of the watermarked speech depends on the relative energy of the embedded watermark, the watermark sequence used, and the LP model order. Robustness of the technique to various signal processing operations and attacks like compression, cropping, and additive noise are studied via experiments on a small order. Robustness of the technique to various signal processing operations and attacks like compression, cropping, and additive noise are studied via experiments on a small order. Robustness of the technique to various signal processing operations and attacks like compression, cropping, and additive noise are studied via experiments on a small order. Robustness of the technique to various signal processing operations and attacks like compression, cropping, and additive noise are studied via experiments on a small order.

1. INTRODUCTION

Digital media and global access to high-speed computer networks are creating complex copyright questions and concerns for legal scholars, librarians, broadcasters, network service providers, publishers, and many others with an interest in legally-protected materials [1]. One response to the unprecedented need for innovation in protecting intellectual property has been the emergence of an active research effort to develop digital “watermarking” strategies. Watermarking generally refers to the process of embedding imperceptible data (the watermark) into a host signal (the coversignal). The watermark might be a pseudo-noise sequence, a sequence of symbols mapped from a message, or linear-prediction coefficients from speech identifying the owner. An important use of watermarking is to offer copyright protection by providing identifying information which, in principle, is known only to, and can be accessed only by, the rightful owner of the material. Only the watermarked version of the material (the stegosignal) is ever released to the public. When copyright questions arise, the watermark is recovered from the stegosignal as evidence of title. Watermarking has been argued to be an advantageous solution to this modern copyright problem, and there is strong evidence that the practice will be accepted by the courts as proof of title [1].

The design of a watermarking strategy involves the balancing of two principal criteria. First, embedded watermarks must be imperceptible to the listener (or viewer). Second, watermarks must be robust. That is, they must be able to survive attacks—those deliberately designed to destroy or remove them, as well as distortions inadvertently imposed upon the watermarks by specific technical processes (e.g., compression) or by systemic processes like channel noise or computational roundoff errors. These two criteria are generally competing in the sense that greater robustness requires more watermark energy, more manipulation of the coversignal, etc., which, in turn, ultimately lead to noticeable distortion of the original content. Related measures of a watermark’s efficacy include data payload, which refers to the number of watermark bits within a unit of time or work [2]. Different watermarking applications have different payload requirements. Another important requirement of a watermarking system is its security, the inherent ability of the system to prevent unauthorized removal, embedding or detection [2]. A watermarking scheme generally derives its security from secret codes or patterns (called keys) that are used to embed the watermark. Only a breach of keying strategies should compromise the security of a watermarking technique; public knowledge of the technical method should not lessen its effectiveness.

The speech watermarking technique described in this paper involves informed embedding, meaning simply that the coversignal is required for watermark recovery. Because of the additional information available during watermark detection and recovery, informed embedding techniques outperform blind watermarking techniques in which the unwatermarked original is not required [2]. Another advantage of informed embedding is that the coversignal serves as a registration mark to undo any temporal or geometric distortions of the stegosignal.

Watermark embedding techniques vary widely in method and purpose. Watermarks may be additive, multiplicative or quantization-based and may be embedded in the space or time domain, or in some transform domain. Each technical variation tends to be more robust to some forms of attacks than to others, and for this and other application-specific reasons, particular strategies may be better-suited to certain tasks. The application that motivated the present work is the creation of the National Gallery of the Spoken Word.

This work was supported by the National Science Foundation of the U.S. under Cooperative Agreement NSF-9817485. Opinions expressed are those of the authors and do not necessarily reflect the views of the NSF.
Let \( \{s[n]\}_{n=\infty}^{\infty} \) denote a coversignal, and let \( \{s[n]\}_{n>k}^{n} \) be the \( k \)th of \( K \) speech frames to be watermarked. Then:

For \( k = 1, 2, \ldots, K \):

1. Using the “autocorrelation method” (e.g., [5, Ch. 5]), derive a set of LP coefficients of order \( M \), say \( \{a_i\}_{i=1}^{M} \), for the given frame.

2. Use the LP parameters in an inverse filter configuration (e.g., [5, Ch. 5]) to obtain the prediction residual on the frame, \( \{e[n] = s[n] - \sum_{i=1}^{M} a_i s[n - i]\}_{n=k}^{\infty} \).

3. Modify the LP parameters in some predetermined way [in the present work, this is done by altering the accompanying autocorrelation sequence] to produce a new set, say \( \{\tilde{a}_i\}_{i=1}^{M} \). The modifications to the LP parameters (or, equivalently in the present case, to the autocorrelation sequence) comprise the watermark.

4. Use the modified parameters as a (suboptimal) predictor of the original sequence, adding the residual obtained in Step 2 above at each \( n \), to resynthesize the speech over the frame, \( \{\tilde{s}[n] = \sum_{i=1}^{M} \tilde{a}_i s[n - i] + e[n]\}_{n=k}^{\infty} \).

5. Replace \( \{s[n]\}_{n=k}^{n} \) by \( \{s[n]\}_{n=k}^{n} \) in the coversignal.

Next \( k \).

Table 1. Watermark Embedding Algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Using the “autocorrelation method” (e.g., [5, Ch. 5]), derive a set of LP coefficients of order ( M ), say ( {a_i}_{i=1}^{M} ), for the given frame.</td>
</tr>
<tr>
<td>2</td>
<td>Use the LP parameters in an inverse filter configuration (e.g., [5, Ch. 5]) to obtain the prediction residual on the frame, ( {e[n] = s[n] - \sum_{i=1}^{M} a_i s[n - i]}_{n=k}^{\infty} ).</td>
</tr>
<tr>
<td>3</td>
<td>Modify the LP parameters in some predetermined way [in the present work, this is done by altering the accompanying autocorrelation sequence] to produce a new set, say ( {\tilde{a}<em>i}</em>{i=1}^{M} ). The modifications to the LP parameters (or, equivalently in the present case, to the autocorrelation sequence) comprise the watermark.</td>
</tr>
<tr>
<td>4</td>
<td>Use the modified parameters as a (suboptimal) predictor of the original sequence, adding the residual obtained in Step 2 above at each ( n ), to resynthesize the speech over the frame, ( {\tilde{s}[n] = \sum_{i=1}^{M} \tilde{a}<em>i s[n - i] + e[n]}</em>{n=k}^{\infty} ).</td>
</tr>
<tr>
<td>5</td>
<td>Replace ( {s[n]}<em>{n=k}^{n} ) by ( {s[n]}</em>{n=k}^{n} ) in the coversignal.</td>
</tr>
</tbody>
</table>

(NGSW), a Digital Libraries Initiative II project whose goal is the development of a carefully organized on-line repository of spoken word collections, based largely upon the renowned Vincent Voice Library at Michigan State University (MSU). Speech engineering aspects of the NGSW are being researched by Speech Processing Laboratory at MSU in collaboration with the Center for Spoken Language Research (CSLR) at the University of Colorado. A brief introduction to the NGSW project by CSLR researchers is found in the ICSLP 2000 proceedings [3], and further information is available at www.ngsw.org and at cslr.colorado.edu.

The method proposed in the present paper is motivated by properties of the special type of signal being watermarked, viz. speech, and, although the technique presented here is readily implemented in real-time, the NGSW application places few constraints on computational load since water-marking can be done off-line if necessary. Further, since the NGSW is a permanent, large-scale, repository of speech data with a rich meta-data support structure, the association of relatively detailed watermarking information with records in the database is not prohibitive. On the other hand, the NGSW material will be compressed for digital streaming, and compression algorithms are notoriously destructive to many types of watermarks. Owners of copyrighted materials are often reluctant to grant permission to post such material on the internet without sufficient assurances that their rights will be protected. Accordingly, a prime interest in the development of the watermarking scheme presented here was the need for robustness to the broadest possible array of attacks. Of course, preserving the audio history and authenticity of the NGSW materials requires that robustness not come at the expense of perceptible distortion.

2. PARAMETER-EMBEDDED WATERMARKS

2.1. Embedding Algorithm

Of the classes of techniques noted above, the watermarking strategy developed in this work is most aptly called a transform domain technique, although watermarks are not embedded in a classical transform domain. The watermarking algorithm presented here is based on the embedding of information (in principle) into the parameters of a signal model for some segment(s) of the coversignal. Embedding information in a parameter domain generally implies the alteration of signal properties that are not linearly related to the signal samples. This renders the embedded information more difficult to separate from the signal, thus offering higher security. This enhanced security, however, need not come at the expense of excessive computational complexity, nor of greater vulnerability to attack. For speech signals, a parametric approach is naturally motivated by the extraordinary successes in applying linear parametric models – in particular, the linear prediction (LP) model (e.g., [5, Ch. 5]) – in several key speech technology areas. The robustness of the LP model to practical anomalies occurring in coding, recognition, and other applications, suggested that some representation of these parameters might provide an effective basis for embedding durable watermarking data. Further adding to the appeal of this embedding venue are the well-understood properties of the LP (and other parametric model) solution, knowledge which is amenable to understanding performance of a watermarking strategy based upon such models.

For simplicity, we base the remainder of this discussion on the LP model of speech. These ideas are readily generalized in interesting and useful ways, and such generalizations will be reported in future work. In general terms, the steps of a (LP) parameter-embedded watermarking procedure are given in Table 1. In anticipation of a more general class of techniques, the algorithm is stated, implicitly at least, in terms of direct parameter modification. In the present work, however, these parameter modifications were made “indirectly” through changes to the autocorrelation sequence – in particular, by adding the watermark sequence pointwise to the autocorrelation. This procedure, which is illustrated in Section 3.1 below, was found to provide a more robust...
watermark than that resulting from direct manipulation of parameters. We note that a number of additional security features are inherent in this approach. These are detailed in Section 3.3.

2.2. Recovery Algorithm

The algorithm for recovering the watermark from the stegosignal (for the special “autocorrelation embedding” procedure employed here) is given in Table 2.

3. APPLICATION AND DISCUSSION

3.1. Example Application

The following example is used as a basis for discussion of the parameter-watermarking technique. The familiar utterance from the TIMIT database [7] “She had your dark suit in greasy wash water all year” was used as the coversignal. The talker is female and the data are sampled at 16 kHz. A single watermark was embedded in the 3001 samples (0.1876 seconds) of speech (corresponding to the utterance “dark”) shown in Fig. 1(a).

![Coversignal](image1.png)

![Stegosignal](image2.png)

Fig. 1. (a) Frame of the coversignal to be watermarked in the example application of Section 3.1. (b) Corresponding frame of stegosignal.

An $M = 18$ order LP model for the speech frame, $\{a_i\}$, was obtained via the Levinson-Durbin recursion (“autocorrelation method” [5, Ch. 5]), and the residual prediction error sequence, $\{e[n]\}$, was computed. Model order $M = 18$ was selected because it provided a good trade-off between robustness and fidelity constraints. The “true” LP model coefficients were altered indirectly by adding the sequence (nominal watermark) $\{0, 0, 0, 0, 2, 7, 5, 2, 7, 6, 3, 6, 2, 3, 7, 9, 4, 0, 0\} \times 0.01$ to the accompanying autocorrelation sequence. The factor 0.01 scaled down watermark energy to ensure watermark imperceptibility. The watermarked speech frame was then created according to Step 4 in Table 1. The resulting stegosignal is shown in Fig. 1(b).

The algorithm given in Table 2 was used to recover the watermark from the stegosignal. All watermark values were found to be accurate to 13 decimal places with respect to the integer sequence indicated above.

3.2. Robustness Issues

Robustness of the foregoing methods to various attacks is discussed in this section. The experiments were conducted for the stegosignal constructed in Section 3.1. Space does not permit a thorough discussion of results which will appear in a more comprehensive account of this work.

Robustness is the ability of the watermark to survive distortion or a signal processing operation to the extent that the quality of the coversignal is not affected beyond a set standard, or that the watermark detection and recovery processes are not hindered. Some of the factors affecting the robustness of the present technique include length of the speech frame to be watermarked, the watermark sequence, the relative energy of the watermark, the LP model order, and the locations of the watermarks in the stegosignal.

**Robustness to additive noise.** Uncorrelated additive noise (Gaussian or uniform distribution) was added to the stegosignal. Table 3 shows miss rates (the recovered watermark differed from the original watermark beyond round-off errors) for various signal-to-noise ratios (SNRs). The miss rates were obtained by repeating the experiment 1000 times for every specification of the noise density parameters. The robustness can be significantly improved if the embedded watermark sequences have inherent error correction capability [2].

**Robustness to MPEG 1 layer III (mp3) compression.** The parametric model based watermarking technique was robust to mp3 compression. Experiments were conducted for a wide range of constant bit rates (CBR) and variable bit rates (VBR). The watermark was recovered efficiently even for very low bit rates like 8 kbps CBR or 10kbps VBR.

**Robustness to cropping.** In a cropping attack, arbitrary samples of the stegosignal are removed. Since the parametric modeling based watermarking involves an additive operation during the watermark embedding and recovery processes, cropping results in desynchronization of the coversignal and the stegosignal resulting in the watermark not being recovered correctly. However, as the present method is an informed watermarking technique, the algorithm described in [8] can be used for resynchronization of the cover and stegosignals facilitating watermark recovery.

**Robustness to jitter attack.** To implement a jitter attack, arbitrary samples of the stegosignal were duplicated. The speech watermarking technique was sufficiently robust against jitter attack even when one in every 12 samples was duplicated. In a modified implementation of the jitter attack that involved zeroing arbitrary samples of the stegosignal, the technique was robust when one out of every 30 samples was set to zero.
Table 2. Watermark Recovery Algorithm (special case of parameter modification by autocorrelation embedding)

For $k = 1, 2, \ldots, K$

1. Subtract residual frame $\{e[n]\}_{n=1}^N$ from the stegosignal frame $\{s[n]\}_{n=1}^N$. This results in an estimate of the modified predicted speech, $\{\hat{s}[n] = \hat{s}[n] - e[n]\}_{n=1}^N$.

2. Estimate the modified LP coefficients $\{\hat{a}_i\}_{i=1}^M$ by computing the least-square-error solution, say $\{\hat{a}_i\}_{i=1}^M$, to the overdetermined system of equations: $\{\hat{a}_i\}_{i=1}^M = \arg\min_\alpha \sum_{n=1}^N \left[\hat{s}[n] - \sum_{i=1}^M a_i s[n - i]\right]^2$.

3. Use the parameter estimates from Step 2 to derive the corresponding autocorrelation values (e.g., see [6, pp. 261-3]). The estimate of the embedded watermark is obtained by subtracting the original autocorrelation values for the $k$th frame of the coversignal from the estimated modified autocorrelation values.

Next $k$.

Robustness to requantization. The speech watermarking technique was fairly robust against requantization. The stegosignal, initially quantized using 16 bits was requantized using 8 bits. The correlation coefficient between the original watermark and the recovered watermark on requantization was 0.9991.

Table 3. Robustness to Uncorr’d Additive Noise

<table>
<thead>
<tr>
<th>Noise distrib’ns</th>
<th>SNR range (dB)</th>
<th>Correl’n coef.* (range)</th>
<th>Wmk. miss rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>39.1 - 39.6</td>
<td>0.9999 - 0.9998</td>
<td>0.0</td>
</tr>
<tr>
<td>uniform</td>
<td>37.2 - 37.6</td>
<td>0.999 - 0.995</td>
<td>0.1</td>
</tr>
<tr>
<td>uniform</td>
<td>35.6 - 36.1</td>
<td>0.999 - 0.993</td>
<td>0.3</td>
</tr>
<tr>
<td>Gaussian</td>
<td>36.1 - 36.9</td>
<td>0.999 - 0.995</td>
<td>0.0</td>
</tr>
<tr>
<td>Gaussian</td>
<td>34.4 - 35.2</td>
<td>0.999 - 0.993</td>
<td>0.4</td>
</tr>
<tr>
<td>Gaussian</td>
<td>33.1 - 33.8</td>
<td>0.999 - 0.988</td>
<td>2.0</td>
</tr>
</tbody>
</table>

*Correl’n coef. between true and recovered watermarks

Robustness to filtering. The technique was not, however, sufficiently robust against filtering in its present formulation. This anomaly remains an issue for further investigation. A possible solution to this problem might be to add the watermark to autocorrelation values based on selective frequency regions of the coversignal.

3.3. Security Issues

To summarize points made above, security refers to a watermark’s ability to withstand attacks aimed mainly at unauthorized removal, detection, or embedding. A watermark technique must not rely on the secrecy of the algorithm for its security. The present watermarking technique relies on the following aspects for its security: Use of pseudo-random watermark patterns (cipher keys), speech frames to be watermarked can be selected randomly, and the LP model order can be different for each watermarked frame of the coversignal (model order also depends on the fidelity constraint). In addition to the above factors, a copy of the coversignal is required for watermark recovery. Because the prediction residual associated with the coversignal is used for reconstructing the stegosignal, the autocorrelation values of the stegosignal are different from the modified autocorrelation values obtained on adding the watermark to the autocorrelation sequence of the coversignal. Hence watermark recovery is rendered very difficult or impossible without a copy of the coversignal.

4. CONCLUSIONS

This paper presents a special instance of a general approach to watermarking of speech signals based on parametric modeling. In experiments presented here, and in many others, the method has proven to be very robust to most common forms of attack, the exception being the filtering attack. Filtering robustness can almost certainly be improved by frequency-selective embedding of the watermark. The method also has many technical features that provide highly-secure watermarking.

5. REFERENCES