

Audio Steganalysis With Content-Independent Distortion Measures

İsmail Avcıbaşı, *Member, IEEE*

Abstract—We first propose a novel content-independent distortion measurement method and use this methodology for digital audio steganalysis. Content-independent distortion measures are utilized as features for the classifier (steganalyzer) design. Experimental results show that the removal of content dependency from features enhances their discriminatory power.

Index Terms—Distortion measures, linear regression, steganalysis, steganography, watermarking.

I. INTRODUCTION

STEGANOGRAPHY is the art and science of hiding the very presence of communication by embedding secret messages into innocent-looking electronic signals such as digital images, video, and audio. To achieve covert communication, *stego-signals*, which are signals containing a secret message, should be indistinguishable from *cover-signals*, which are signals not containing any secret message. In this respect, *steganalysis* is the set of techniques that aim to distinguish between cover-signals and stego-signals. Steganalysis can be implemented in either a *passive warden* or *active warden* style [1]. A passive warden simply examines the signal and tries to determine if it potentially contains a hidden message. If it appears that it does, then the signal is stopped; otherwise, it will go through. An active warden, on the other hand, can alter signals intentionally, even though there may not be any trace of a hidden message, in order to foil any secret communication that nevertheless can be occurring.

While there has been quite some effort in the steganalysis of digital images, and [2] and [3] are good survey papers, steganalysis of digital audio is relatively unexplored. Westfeld and Pfitzmann proposed a steganalysis method [4] for LSB-based embedding. In another paper, Westfeld [5] addressed the steganalysis of the MP3Stego algorithm. Johnson *et al.* proposed the steganalysis of LSB-based embedding and the Hide4pgp algorithm [6]. Ozer *et al.* [7] proposed a universal audio steganalysis technique that is effective on both watermarking and steganographic data-embedding methods. The basic idea in [7] rests on the statistical evidence that the distortion measures computed between signals and their denoised versions have statistically distinguishable distributions for cover-signals and stego-signals. These statistically distinguishable features are used in steganalyzer design to classify cover-signals from stego-signals.

Manuscript received May 31, 2005; revised November 2, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mauro Barni.

The author is with the Department of Electronics Engineering, Uludag University, 16059 Bursa, Turkey (e-mail: avcibas@uludag.edu.tr).

Digital Object Identifier 10.1109/LSP.2005.862152

In this letter, we first analyze the inherent content dependency in distortion measures as calculated in [7]. Second, we propose a novel method to remove this content dependency from distortion measurements. By a statistical hypothesis test, we justify how the proposed technique enhances the discriminatory power of the features used in the classifier. These content-independent measurements are then used as features to build a classifier to differentiate cover-signals and stego-signals.

The rest of the letter is organized as follows: In Section II, we describe how to remove content dependency from distortion measurements. Steganalyzer design with content independent distortion measurements and results are given in Sections III and IV. Conclusions are drawn in Section V.

II. CONTENT-INDEPENDENT DISTORTION MEASUREMENT

The potential of certain audio quality metrics in predicting the presence of watermarking and steganographic signals within an audio is shown in [7]. There exists a rationale to utilize more than one distortion measure, in order to probe different quality aspects of the signal, which could be impacted during data-hiding manipulations. In such a task, there is the risk that the variability in the signal content itself eclipses the detector from the alterations. Thus, it is desired that, whatever features are selected, the detector responds only to the *induced distortions* during data hiding and not be confused by the statistics of the signal content. Finally, the original signal (ground-truth) obviously will not be available during the testing stage. Therefore, some “ground-truth” or reference signal must be created that is common to both the training and testing stages. In [7], a denoised version of the given signal is used as the ground-truth reference. However, this self-referencing, which is creating a reference signal via its own denoised version, is obviously a content-dependent scheme. To avoid content dependency, we propose to use a single reference signal that is common to all signals to be tested. Thus, we use a reference signal and its altered versions according to the type of data embedding.

More specifically, let x denote a test signal and $x + \varepsilon$ be its embedded version, and similarly, let y and $y + \eta$ indicate the reference signal and its embedded version. Furthermore, let us consider a generic distortion functional $M(a, b)$ between the signals a and b . For example, for the mean-square distortion, one simply has $M(a, b) \triangleq E[(a - b)^2]$, with E being the expectation operator. The detector is based on the statistics of the difference of the distortions, as will be explained in the sequel. We have, however, two assumptions for the operation of the detector. First, data embedding leads to additive distortion, that is, the altered signals can be represented as $x + \varepsilon$ and $y + \eta$. Second, the additive distortions of the test and reference signals should

not be mutually orthogonal, that is, $E\{\varepsilon^*\eta\} \neq 0$. This assumption was indirectly justified by analysis of variance (ANOVA) and the test results given in the experimental results section.

We first show that self-referencing, as employed in [7], causes content-dependent distortion. Let f be the specific operation to obtain the reference signal; for example, in [7], denoising operation has been used: $y = f(x) = \text{denoise}(x)$. The outcomes of this operation are given by $x \xrightarrow{f} f(x)$ and $x + \varepsilon \xrightarrow{f} f(x + \varepsilon)$, respectively, for original signal and its embedded version. To illustrate the point, for the case of the mean-square distortion, one obtains

$$M(x + \varepsilon, f(x + \varepsilon)) - M(x, f(x)) = E[f(x + \varepsilon)^2 + 2x\varepsilon + \varepsilon^2 - 2(x + \varepsilon)f(x + \varepsilon) + 2xf(x) - f(x)^2] \quad (1)$$

which is content dependent, because the signal terms x and $f(x)$ survive in the difference of distortion functionals. For content independence, the above difference should be some function of only the distortion term ε and should not contain x or any of signal derived from it.

Now we take a different route and take as a reference a unique signal y . We then measure the distortions between x and $x + \varepsilon$, using y and $y + \eta$ as reference signals, where $y + \eta$ represents the embedded version of the reference signal. The relationship of these signals and the distortion *vis-à-vis* the reference signals y and $y + \eta$ is illustrated in Fig. 1. In this figure, the length of the vector $\vec{x}\vec{y}$ is simply equal to $M(x, y)$. The distance between the tips of the vectors $\vec{x}\vec{y}$ and $\vec{x}\vec{(y+\eta)}$ is $d = M(x, y) - M(x, y + \eta)$, and similarly, $d' = M(x + \varepsilon, y) - M(x + \varepsilon, y + \eta)$ denotes the distance between the tips of the dashed pair of vectors. For the case of the mean-square distortion, it follows that $d = E[(x - y)^2 - (x - y)^2 + 2(x - y)\eta - \eta^2] = E[2(x - y)\eta - \eta^2]$ and $d' = E[(x + \varepsilon - y)^2 - (x + \varepsilon - y)^2 + 2(x + \varepsilon - y)\eta - \eta^2] = E[2(x - y)\eta + 2\eta^*\varepsilon - \eta^2]$. Now if one considers the difference of d and of d' , one can observe that one achieves content independence, that is

$$D_1 \triangleq d' - d = 2E[\eta^*\varepsilon]. \quad (2)$$

Let us consider another measure, the correlation coefficient, given by $M(a, b) \triangleq E[ab]$. One can easily show that $d = E[xy] - E[x(y + \eta)] = -E[x\eta]$ and $d' = E[(x + \varepsilon)y] - E[(x + \varepsilon)(y + \eta)] = -E[x\eta] - E[\varepsilon^*\eta]$, so that $D_2 \triangleq d' - d = -E[\varepsilon^*\eta]$. Again the difference of distortions is not a function of image content x and y but rather of the product of distortions $\varepsilon^*\eta$.

III. DESIGN OF THE STEGANALYZER

The methodology presented in Section II is used with a subset of metrics to obtain feature vectors for the steganalyzer design. The audio quality measures (AQMs) (see [7] for details) are classified as follows.

Time domain measures: Signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SNRseg), Czenakowski distance (CZD).

Frequency-domain measures: Log-likelihood ratio (LLR), log-area ratio (LAR), Itakura distance (ID), Itakura-Saito distance (ISD), COSH distance (COSH), Cepstral distance (CD),

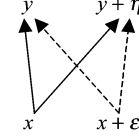


Fig. 1. Configuration of the signal vectors: the original signal x , its embedded version $x + \varepsilon$, the reference signal $y + \eta$, and its embedded version.

short-time Fourier–Radon transform distance (STFRT), spectral phase distortion (SP), spectral phase-magnitude distortion (SPM).

Perceptual-domain measures: Bark spectral distortion (BSD), modified bark spectral distortion (MBSD), enhanced modified bark spectral distortion (EMBSD), perceptual speech quality measure (PSQM), perceptual audio quality measure (PAQM), measuring normalizing block 1 (MNB1), measuring normalizing block 2 (MNB2), weighted slope spectral distance (WSS).

We used a training set of original audio signals and their embedded versions, as well as the original and embedded versions of the reference audio signals. We have used a randomly selected reference audio signal to be used both in the design and in the test. The selection of the reference audio signal is arbitrary; as it is shown in the previous section, measures do not have content dependency. A linear regression classifier was designed using the statistics collected with the database of audio signals.

We use cover $C = \{x_i\}$ and stego $S = \{x_i + \varepsilon_i\}$ sets in the training. Here $i = 1, \dots, N$, and $N = 50$ is number of samples in the training set. By using the methodology as explained in the Section II, for every sample in the training set, we have obtained the feature values (distortion measures) using reference signal pairs y and $y + \eta$. We constructed 50 feature vectors from cover-set C and 50 feature vectors from stego-set S . Then, these feature vectors were regressed to -1 and 1 , depending upon whether the feature vector comes from C and S , respectively. In the regression model [8], we expressed each decision label $g_i \in [-1, 1]$, $i = 1, \dots, N$ as a linear combination of content-independent distortion measures $g_i \in \beta_1 f_{1i} + \beta_2 f_{2i} + \dots + \beta_q f_{qi}$, where $\mathbf{f}_i = (f_{1i}, f_{2i}, \dots, f_{qi})$ is the vector of q content-independent distortion measures computed from the i th audio sample, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)$ are the regression coefficients. Once the regression coefficients were predicted in the training phase, then they were used in the testing phase. In the testing phase, we first constructed the feature vector \mathbf{f} for an incoming audio signal by using the reference cover y and stego $y + \eta$ pair that was used in training phase. Next, $g = \beta_1 f_1 + \beta_2 f_2 + \dots + \beta_q f_q$ is evaluated. If the evaluated value exceeds the threshold 0, then the decision is that the incoming audio signal contains message; otherwise, the decision is that it does not.

IV. EXPERIMENTAL RESULTS

Steganographic algorithms differ in hiding the messages. Least-significant bit (LSB) methods embed the message by flipping the LSBs of audio samples [9], [10] or, alternatively, transform coefficients [11], [12]. Spread-spectrum techniques add a scaled and spread version of the message into the cover signal in the time or frequency domain, possibly with perceptual weighting to guaranty inaudibility [12]. A secret message

TABLE I
PERFORMANCE OF THE CLASSIFIERS

Data Hiding Methods	False Positive		False Negative	
	AQM	CIAQM	AQM	CIAQM
DSSS	0/50	0/50	0/50	0/50
FHSS	0/50	0/50	1/50	0/50
ECHO	1/50	0/50	0/50	0/50
STEGANOS	6/50	3/50	0/50	1/50
HIDE4PGP	14/50	4/50	11/50	8/50
STEGHIDE	13/50	8/50	14/50	6/50

TABLE II
FEATURES USED IN STEGANALYZER DESIGNS

	AQM	CIAQM
DSSS	SNRs, LLR, LAR	LLR, ISD, COSH
FHSS	SNRs, LLR, LAR, CDM, WSSD, PAQM, CZD, SP, STFRT	LLR, ISD, COSH
ECHO	SNRs, LLR, MBSL, WSSD, PAQM, CZD, SP, STFRT	LLR, ISD, COSH, SP
STEGANOS	PAQM, CZD, STFRT	LLR, ISD, COSH, SNR
HIDE4PGP	LAR, COSH, EBSD, PAQM, CZD	LLR, ISD, COSH, SP
STEGHIDE	SNRs, LLR, EBSD, PAQM, CZD, SP	LLR, ISD, COSH, CZD

can also be inserted into an audio signal during the quantization and dithering process [13].

We have performed steganalysis experiments over six different algorithms, three of which are watermarking techniques and three of which are steganographic techniques. The watermarking techniques are direct-sequence spread spectrum (DSSS) [12], frequency hopping with spread spectrum (FHSS) [12], and echo hiding (ECHO) [12]. The steganographic methods are Steghide [14], Hide4pgp [15], and Steganos [10]. The rationale of using these tools was their popularity, high embedding capacity, free availability, wide usage, and applicability to audio signals. It may be questionable to use watermarking techniques within the context of steganalysis since watermarking does not attempt to be undetectable. However, in the active warden scenario, data must be embedded into the cover robustly so that the stego-signal may still carry the message after the warden takes deliberate actions on the signal.

We use recorded speech segments for the experimentation. The speech segments have durations of three to four seconds, are sampled at 16 kHz, and are recorded in an acoustically shielded medium. The procedure consists of embedding messages to all available cover signals, 100 speech utterances, randomly selecting half of the set of the stego and cover signals for training, and leaving the other 50% for testing phase. The embedded message size was around 10% of the audio size, which is usually the maximum allowed capacity for LSB embedding. Since the sample size was small, we limited the number of used features to 5, which is the number corresponding to 10% of the number of training samples. As a rule of thumb, this number of features is assumed to be optimum for the classifier to generalize [16].

The detection results by the proposed content-independent audio quality metrics (CIAQMs) are compared to [7] in Table I. In Table II, we give the metrics' names used in the design of steganalyzers to obtain the results given in Table I. The classifier used in each scheme was the linear regression classifier on the same data set. From the results, we see that AQM and CIAQM methods are equally quite effective on watermarking methods. However, the CIAQM method is more effective on

TABLE III
 p -VALUES OF ANOVA TEST FOR LLR DISTORTION MEASURE

p -value for LLR	DSSS	FHSS	ECHO	STEGA NOS	HIDE4 PGP	STEG HIDE
AQM	0.0000	0.0002	0.0001	0.0952	0.1118	0.0318
CIAQM	0.0000	0.0000	0.0000	0.0004	0.0002	0.0015

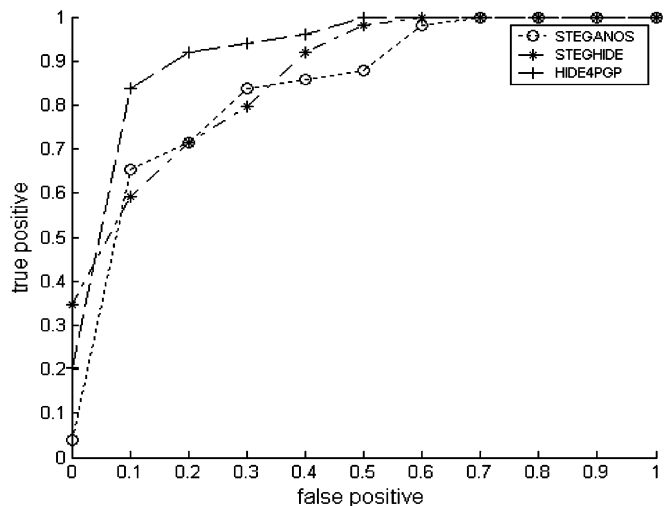


Fig. 2. ROC curves of Steganos, Steghide, and Hide4Pgp for the proposed CIAQM method.

steganographic data-embedding methods than the AQM method is. It should be noted that, even if not audible, the incurred distortion in watermarking is more pronounced than the distortion incurred in steganographic data embedding. This effectiveness can be attributed to the removal of content dependency from the distortion measurements.

The success of a classifier highly depends on the features' discriminatory power. To see how the proposed technique enhances the discriminatory power of the features, we have applied the ANOVA test to the features to determine if the variations of a measure result from the content of the cover signal or the presence of a hidden message. ANOVA is a general statistical hypothesis testing technique used when one wants to determine whether or not a number of data groups are statistically different [8]. The p -value is the probability of finding in reality that there is no difference between the means. Thus, in our case, the lower the p -value, the more discriminative the feature. Table III gives the p -values of the ANOVA test for the null hypothesis that the means of the groups are equal for the chosen LLR distortion measure on the same data set. We have chosen this measure as it is a common feature used in both schemes that are compared. We used one-way ANOVA to test the LLR feature for different embedding methods. For a given embedding method, each group consisted of 50 samples of LLR features computed from the cover-signals and stego-signals, respectively, in the training set. Relatively smaller p -values for the LLR measure indicate the relative improvement on the discriminatory power and shed light on the performance improvement for steganographic data embedding. It can be seen that LLR in AQM cannot significantly discriminate for Steganos and Hide4PGP. The features in Table II for CIAQM were selected based on ANOVA. The features with the lowest p -values obtained by ANOVA were used in the design of Steganalyzers. The ANOVA was performed for

each feature using the training samples and optimized on the training set. The features in AQM are selected by the floating search method (FSM) [17]. FSM basically selects the optimum subset of features yielding the minimum error for a classifier (linear regression classifier in our case).

In order to show the relationship between the false-positive rate and the detection rate, we have also calculated the receiver operating characteristics (ROC) curves of steganographic data embedding for the CIAQM method in Fig. 2. The ROC curves are calculated for the linear regression classifiers by first designing a classifier and then testing the data unseen to the classifier against the trained classifier at the same time changing the threshold of decision. As the threshold is changed, the false-positive rate changes, and for each false-positive rate, we get a corresponding detection rate.

V. CONCLUSION

In this letter, a methodology was presented for content-independent distortion measurement. This methodology was then used for audio steganalysis, where content-independent distortion measures were used as features in the design of linear regression classifier. The steganalyzer was tested on watermarking and steganographic methods. Removal of content dependency from the measurements enhanced their discriminatory power and proved to be useful, especially for steganographic data embedding, where the incurred distortions are much less pronounced than in watermarking.

ACKNOWLEDGMENT

The author would like to thank H. Ozer of the TUBITAK UEKAE Speech Group for providing the database of digital speech tracks used in the experiments.

REFERENCES

- [1] G. J. Simmons, "Prisoners' problem and the subliminal channel," in *Proc. CRYPTO83—Advances Cryptology*, 1984, pp. 51–67.
- [2] J. Fridrich and M. Goljan, "Practical steganalysis of digital images—State of the art," in *Proc. SPIE Photonics West*, vol. 4675, 2002, pp. 1–13.
- [3] R. Chandramouli, M. Kharrazi, and N. D. Memon, "Image steganography and steganalysis: Concepts and practice," in *Proc. IWDW*, 2003, pp. 35–49.
- [4] A. Westfeld, "Detecting low embedding rates," in *Proc. 5th Int. Workshop, Information Hiding*, F. A. P. Petitcolas, Ed., Noordwijkerhout, The Netherlands, Oct. 7–9, 2002, pp. 324–339.
- [5] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," in *Information Hiding*, ser. Lecture Notes in Computer Science. Heidelberg, Germany: Springer-Verlag, 1999, vol. 1768, pp. 61–66.
- [6] M. K. Johnson, S. Lyu, and H. Farid, "Steganalysis of recorded speech," in *Proc. SPIE*, vol. 5681, Mar. 2005, pp. 664–672.
- [7] H. Ozer, I. Avcibas, B. Sankur, and N. Memon, "Steganalysis of audio based on audio quality metrics," in *Proc. SPIE Security Watermarking Multimedia Contents V*, vol. 5020, 2003, pp. 55–66.
- [8] C. Rencher, *Methods of Multivariate Analysis*. New York: Wiley, 1995.
- [9] Stools and A. Brown. (1996) S-Tools Version 4.0, Copyright C. [Online]. Available: <http://members.tripod.com/steganography/stego/s-tools4.html>.
- [10] Steganos [Online]. Available: <http://www.steganos.com>.
- [11] I. Cox, J. Kilian, F. T. Leighton, and T. Shamoan, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.
- [12] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.*, vol. 35, no. 3&4, pp. 313–336, 1996.
- [13] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1423–1443, May 2001.
- [14] S. Hetzl. [Online]. Available: <http://steghide.sourceforge.net>.
- [15] H. Repp. (2000) Hide4PGP. [Online]. Available: <http://www.heinz-repp.onlinehome.de/Hide4PGP.htm>.
- [16] A. K. Jain, R. P. W. Duin, and M. Jianchang, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
- [17] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, pp. 1119–1125, 1994.