Statistical Analysis of Two Polarity Detection Schemes in Speech Watermarking

Bin Yan[1], Zhe-Ming Lu[1], Jeng-Shyang Pan[2], Sheng-He Sun[1]

[1]Department of Automatic Test and Control, Harbin Institute of Technology,
P. O. Box 339, 150001 Harbin, P. R. China.
yanbinhit@hotmail.com
zhemingl@yahoo.com
[2]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences,
415 Chien-Kung Road, Kaohsiung 807, Taiwan, R.O.C.
jspan@cc.kuas.edu.tw

Abstract

Polarity inversion based speech watermarking scheme hide data in speech by modification of the speech polarity. This paper build a statistical model of the polarity detection problem, based on this model, the original polarity detection scheme and the optimal detection scheme are analyzed and compared. The theoretical analysis results are validated by Monte Carlo simulation, the optimal polarity detector shows significant performance gain compared with the original polarity detection algorithm.

1 Introduction

Polarity Inversion (PI) based watermarking scheme utilizes the fact that the human auditory system (HAS) is insensitive to the polarity of the speech signal [1]. Secure data can be hidden in speech signal by inverting the polarity of certain portion of the signal. PI watermarking can be classified as phase coding scheme [2, 3], the phase of the speech is changed by 180 degrees for PI watermarking. PI is very robust against noise addition and filtering operations because the polarity of the voiced frame won't change under these manipulations. The drawback of PI watermarking is that it is not secure, but it is very useful for content annotation and in-band signalling applications [4]. The problem to be solved by this paper is to evaluate the performance of the polarity detection algorithms. This is done by first building a statistical model for the speech residual signal, based on this model, the performance of the original polarity detection algorithm is analyzed, this result is compared to the performance of the optimal polarity detector. Finally, we perform Monte Carlo simulation to validate the theoretical analysis.

2 Detection Performance of the Original Method

The polarity detection scheme proposed by [1] can be summarized as a two-step procedure for each syllable, first, the polarity of the maximum peak in the LPC residual signal of each voiced frame is estimated, second, the polarity of each syllable is determined by majority vote. In this section we will analyze the error probability of this detection scheme. When the AR model order is properly chosen, the residual signal of the voiced frame can be modeled as impulse train in Additive White Gaussian Noise (AWGN), so we can build the following model under each hypothesis:

$$\mathcal{H}_0 \quad : \quad s[n] = -\sum_{k=0}^{K-1} A\delta[n-kP] + w[n] \tag{1}$$

$$\mathcal{H}_1 \quad : \quad s[n] = + \sum_{k=0}^{K-1} A\delta[n - kP] + w[n] \ , \tag{2}$$

where P is the pitch period, A is the amplitude of the impulse train, K is the number of pitch period in each frame, $w[n]$ is Gaussian noise with mean zero and variance σ^2. For ease of analysis, we made the following assumptions: the sample values in the location of impulses are not affected by the AWGN, the validity of this assumption will be verified by Monte Carlo simulation. Under this assumption, the probability of detection error in each frame is

$$P_{\mathrm{EF}} = \frac{1}{2} \Pr\left\{\max(\mathbf{s}) > A | \mathcal{H}_0\right\} + \frac{1}{2} \Pr\left\{\min(\mathbf{s}) < -A | \mathcal{H}_1\right\} \ ,$$

where $\mathbf{s} = [s[0], \cdots, s[N-1]]^T$, it is assumed that $P(\mathcal{H}_0) = P(\mathcal{H}_1) = 1/2$. The error probability under $\mathcal{H}_1$ is calculated as

$$\begin{aligned} & \Pr\left\{\min(\mathbf{s}) < -A | \mathcal{H}_1\right\} = \Pr\left\{\min(\mathbf{w}) < -A | \mathcal{H}_1\right\} \\ = \ & 1 - \left[1 - Q\left(\frac{A}{\sigma}\right)\right]^N \ , \end{aligned}$$

where $Q(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) \mathrm{d}t$. By symmetry of the problem and the noise distribution, the error probabilities under each assumption are equal, so we have

$$P_{\mathrm{EF}} = 1 - \left[1 - Q\left(\frac{A}{\sigma}\right)\right]^N \ .$$

Suppose that the voiced portion of one syllable has M frames, since the estimation error probability of each frame is P_{EF}, then the error probability of final decision by majority vote is

$$P_{\mathrm{E}} = \sum_{m=\lceil M/2 \rceil}^{M} \binom{M}{m} P_{\mathrm{EF}}^m \left(1 - P_{\mathrm{EF}}\right)^{M-m} \ , \tag{3}$$

where $\lceil x \rceil$ rounds x to the nearest integers towards $+\infty$.

3 Detection Performance of the Optimal Detector

In this section, we consider a more systematic approach for detecting speech polarity using signal detection framework. Fig. 1 shows the signal generation model for voiced speech, If the information bit to hide is 0, the speech signal is modeled as the output of an all-pole model excited by the summation of $u_0[n]$ and $w[n]$, otherwise, the excitation signal is the summation of $u_1[n]$ and $w[n]$. Let $h[n]$ be the impulse response of the all-pole system, then the detection problem is to distinguish between the following two hypotheses

$$\begin{aligned} \mathcal{H}_0 : x[n] &= -\sum_{l=1}^{P_{\mathrm{AR}}} a_l x[n-l] - \sum_{k=0}^{K-1} A\delta[n - kP] + w[n] \\ &= u_0[n] * h[n] + w[n] * h[n] \\ &= \hat{u}_0[n] + \hat{w}[n] \\ \mathcal{H}_1 : x[n] &= -\sum_{l=1}^{P_{\mathrm{AR}}} a_l x[n-l] + \sum_{k=0}^{K-1} A\delta[n - kP] + w[n] \\ &= u_1[n] * h[n] + w[n] * h[n] \\ &= \hat{u}_1[n] + \hat{w}[n] \end{aligned}$$

The parameters $\{a_l\}_{l=1}^{P_{\mathrm{AR}}}, P_{\mathrm{AR}}, P, K$ are assumed known or can be estimated from the speech signals [5]. Since $\hat{w}[n]$ is the output of an all-pole model excited by IID WGN, so $\hat{w}[n]$ is

stationary WGN with mean zero and covariance matrix $\mathbf{C}$. To minimize the probability of decoding error, the optimal detection statistic for distinguishing between $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_0$ can be found to be [6][7]

$$T(\mathbf{x}) = \mathbf{x}^T \mathbf{C}^{-1} (\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0) \quad , \tag{4}$$

which minimizes the probability of detection error. When N is large and the noise is wide sense stationary(WSS), the test statistic is approximated by

$$T(\mathbf{x}) \simeq \int_{-1/2}^{1/2} \frac{X(f) \left[\hat{U}_1(f) - \hat{U}_0(f)\right]^*}{P_{\hat{w}\hat{w}}(f)} \mathrm{d}f \quad ,$$

where $X(f), \hat{U}_1(f), \hat{U}_0(f)$ are the DTFT of $x[n], \hat{u}_1[n], \hat{u}_0[n]$ respectively. $P_{\hat{w}\hat{w}}(f)$ is the power spectrum density(PSD) of the noise $\hat{w}[n]$, i.e.,

$$P_{\hat{w}\hat{w}}(f) = \frac{\sigma^2}{\left|1 + \sum_{l=1}^{P_{\mathrm{AR}}} a_l \exp(-j2\pi f l)\right|^2} \quad .$$

The test statistic can be further simplified by invoking the Parseval's theorem, so we have

$$\begin{aligned} T(\mathbf{x}) &\simeq \int_{-1/2}^{1/2} \frac{X(f) \left[\hat{U}_1(f) - \hat{U}_0(f)\right]^*}{\sigma^2} \left|1 + \sum_{l=1}^{P_{\mathrm{AR}}} a_l \exp(-j2\pi f l)\right|^2 \mathrm{d}f \\ &= \frac{2}{\sigma^2} \sum_{n=P_{\mathrm{AR}}}^{N-1} \left(x_{\mathrm{w}}[n] \sum_{k=0}^{K-1} A\delta[n - kP] \right) \\ &= \frac{2}{\sigma^2} \sum_{k=0}^{K-1} A x_{\mathrm{w}}[kP] \quad , \end{aligned}$$

where x_{w} is the whitened $x[n]$ by inverse filtering [5].

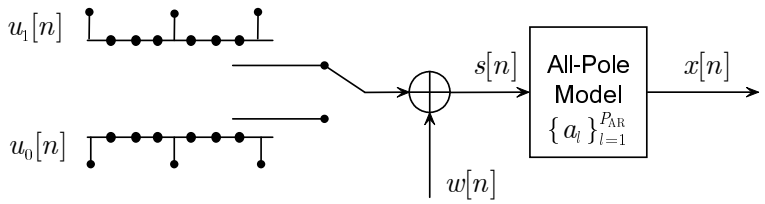

Figure 1: Polarity inversion watermarking model

The detection threshold γ is found to be

$$\gamma = \frac{1}{2} \left(\hat{\mathbf{u}}_1^T \mathbf{C}^{-1} \hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0^T \mathbf{C}^{-1} \hat{\mathbf{u}}_0 \right) \quad .$$

It can be shown that

$$\hat{\mathbf{u}}_1^T \mathbf{C}^{-1} \hat{\mathbf{u}}_1 = \hat{\mathbf{u}}_0^T \mathbf{C}^{-1} \hat{\mathbf{u}}_0 = \frac{A^2 K}{\sigma^2} \quad ,$$

which implies that the detection threshold is $\gamma = 0$. In summary, the optimal detector decide $\mathcal{H}_1$ if

$$T(\mathbf{x}) = \frac{2}{\sigma^2} \sum_{k=0}^{K-1} A x_{\mathrm{w}}[kP] > 0 \quad . \tag{5}$$

The detector performance in terms of probability of error can be proved to be

$$P_{\mathrm{E}} = Q\left[\frac{1}{2} \sqrt{(\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0)^T \mathbf{C}^{-1} (\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0)} \right] \tag{6}$$

$$= Q\left(\sqrt{\frac{A^2 K}{\sigma^2}} \right) \quad . \tag{7}$$

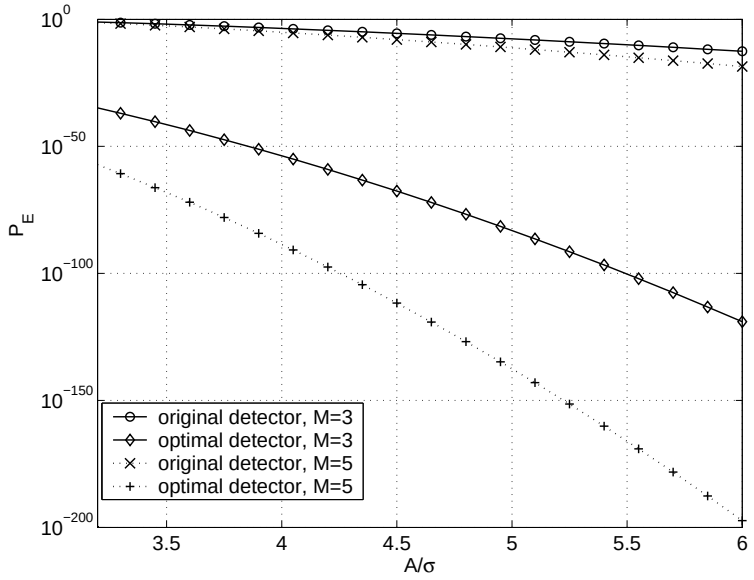

Figure 2: Comparison of theoretical P_{E}

Table 1: Parameters of Monte Carlo simulation to validate the assumptions

Parameter	*Value*
f_s	8kHz
P	60
N	300
A	from 3 to 5
σ	1

Performance Comparison of the Two Methods

In order to compare the theoretical results of (3) and (7), we set $K \times P = N \times M$. The pitch period P is chosen as 60. The results are shown in Fig. 2, the theoretical P_{E} of the optimal detector outperforms the original detector by tens of order of magnitude. When the number of frames in the voiced segment increases, more information-carrying samples are available, the P_{E} of both detectors decrease, this is shown in the figure for $M = 3$ and $M = 5$.

4 Monte Carlo Simulation

In this section, we perform the Monte Carlo Simulation to validate the theoretic analysis.

Validation of the Assumption in Section 2

In section 2, we have made the following assumptions to simplify the analysis: the sample values in the location of impulses are not affected by the AWGN, it is also assumed that the amplitude of the impulse is larger than the maximum absolute value of the AWGN. Here we will use Monte Carlo simulation to evaluate the effects of these assumptions on the final results. The parameters used in the simulation are shown in Table 1. The Monte Carlo simulation results are shown in Fig. 3 for $A = 3, 4, 5$, the comparison between analytical results and simulation results reveals that the analytical P_{EF} is about ten times larger than the simulation results, however, the analytical results with the assumptions provide an upper bound for the true situations, the assumptions tends to be more realistic for larger A/σ. Furthermore, in the above comparison

between the performance of the original and the optimal detector, we see that the P_{E} of the optimal detector is tens of order of magnitude smaller than the non-optimal case, so the analytical results are valid for comparison purpose.

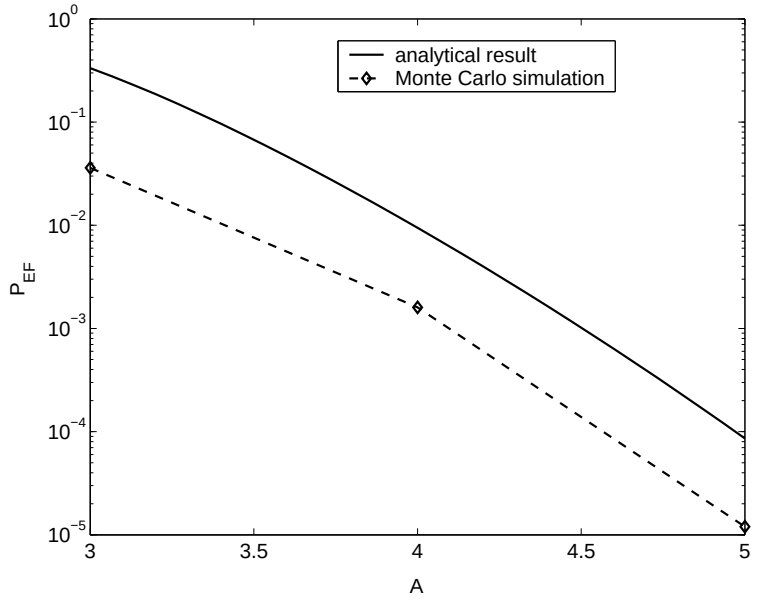

Figure 3: Comparison between analytical results and Monte Carlo simulation results of P_{EF}

Validation of the Impulse Train + WGN Residual Model

In order to validate the "Impulse Train + WGN" model for the residual signal, we perform the following statistical experiments on the real speech signals: first, we compute the residual signal of the voiced frames, then, the impulses are treated as outliers [8], they were eliminated from the residual signal, the histogram of the the remaining residual signal are calculated and compared to the empirical Gaussian PDF with parameters estimated from the data by maximum likelihood (ML) estimation. The outlier detection scheme was based on method proposed in [9], which calculate $M_i = 2\,(x_i - x_{50\%})/(x_{84\%} - x_{16\%})$ for each data sample x_i, where $x_{50\%}$ is the median of data sequence $\{x_i\}_{i=1}^{N}$, $x_{84\%}$ and $x_{16\%}$ are the 84% and 16% percentile respectively, x_i is classified as outlier when $|M_i| > 3$. Fig. 4 shows the experimental results when applying the outlier detection and elimination algorithm on residual signals, due to the non-stationary nature of the speech signal, the outlier detection and elimination algorithm were performed frame by frame . After the removal of outliers, the histogram of the residual signal is fitted by Gaussian distribution. The result is shown in Fig. 5, which shows good matching between the histogram and the Gaussian distribution PDF. The sample mean and standard deviation is estimated to be -0.0016 and 0.0277 respectively. The amplitude of the impulses are found to be between 0.1 to 0.15, the quantity A/σ is between 3 and 6. The pitch period can be found manually to be 40 and 41. Using the parameters estimated above, the synthesized speech residual signal is shown in Fig. 6. We will use this model to validate the theoretical P_{E} of optimal detector by Monte Carlo simulation.

Validation of P_{E} of the Optimal Detector

To validate the results in (7), we perform the Monte Carlo simulation to estimate $\hat{P}_{\mathrm{E}}$. The detector (5) is applied on data sequences generated by the "impulse train + WGN " model, the number of detection errors is counted, the estimated $\hat{P}_{\mathrm{E}}$ can be calculated as

$$\hat{P}_{\mathrm{E}} = \frac{\#\text{ of detection errors}}{\#\text{ of Monte Carlo simulations}} \quad .$$

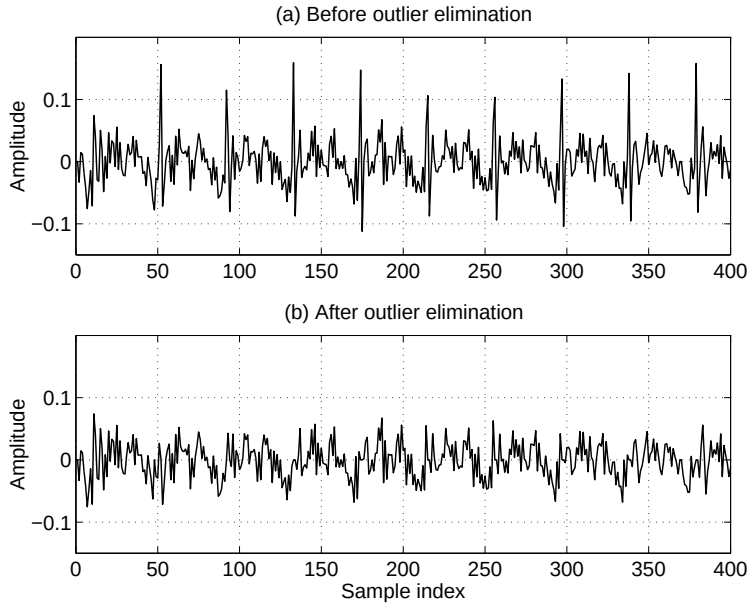

Figure 4: Speech residual signal before and after outlier elimination

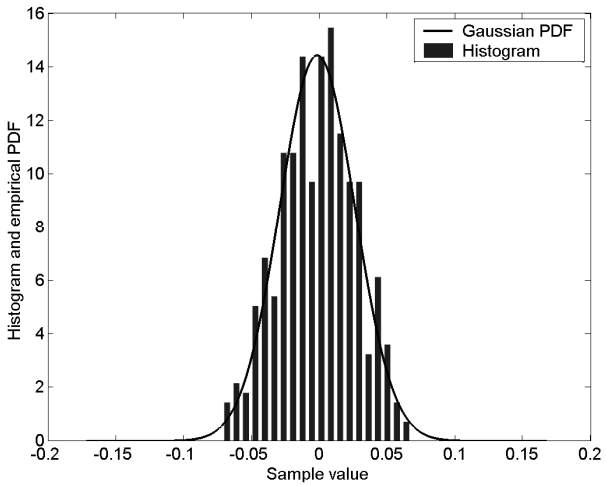

Figure 5: Histogram and Gaussian PDF with parameters estimated from speech residual after elimination of outliers

From the theoretical analysis, we see that the detection error is rather rare event, for example, when $A/\sigma = 5, M = 3$, P_{E} is approximately 10^{-80}, to simulate the rare event, we use importance sampling to reduce the variance of $\hat{P}_{\mathrm{E}}$ [10, 11]. Due to the symmetry of the detection problem and the noise distribution, we only consider the detection error under $\mathcal{H}_1$, which is

$$\begin{aligned} P_{\mathrm{E}} = P_{\mathrm{E}|\mathcal{H}_1} &= \Pr\left\{\frac{1}{K}\sum_{k=0}^{K-1} w[kP] < -A; \mathcal{H}_1\right\} \\ &= \Pr\left\{\frac{1}{K}\sum_{k=0}^{K-1} w[kP] > A; \mathcal{H}_1\right\} = \mathcal{E}_f\left(I_{\left\{K^{-1}\sum_{k=0}^{K-1} w[kP] > A\right\}}\right) \\ &= \mathcal{E}_g\left(I_{\{\bar{w} > A\}}\right) \\ &= \mathcal{E}_{\mathrm{A}}\left\{I_{\{\bar{w} > A\}} \exp\left[\left(-2A\bar{w} + A^2\right) / \left(2\sigma^2/K\right)\right]\right\} \end{aligned}$$

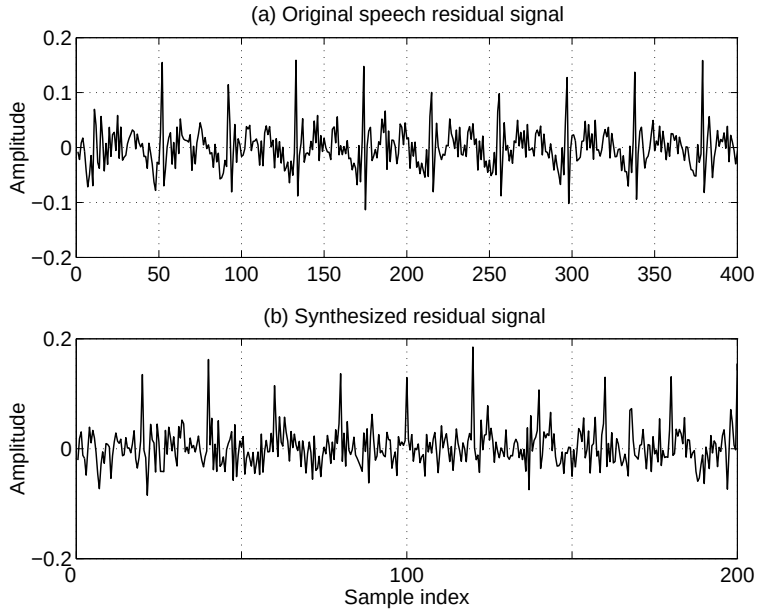

Figure 6: Synthesized speech residual using the "impulse train + WGN " model

Table 2: Theoretical P_{E} and estimated $\hat{P}_{\mathrm{E}}$ by importance sampling

$\frac{A}{\sigma}$	P_{E}	$\hat{P}_{\mathrm{E}}$	95% confidence interval
3	1.6508×10^{-31}	1.8005×10^{-31}	$[1.4065\ 2.1946]\times 10^{-31}$
4	1.9664×10^{-54}	2.0208×10^{-54}	$[1.4851\ 2.5566]\times 10^{-54}$
5	7.6301×10^{-84}	7.0208×10^{-84}	$[4.8998\ 9.1419]\times 10^{-84}$
6	9.4276×10^{-120}	1.1016×10^{-119}	$[0.7480\ 1.4551]\times 10^{-119}$

where $I_{\mathrm{D}}(x)$ is the indicator function, which is one if $x \in D$, and zero otherwise, $\mathcal{E}_f$ is the statistical expectation w.r.t. the distribution $f = \mathcal{N}(0, \sigma^2)$, $\mathcal{E}_g$ is the statistical expectation w.r.t. the distribution $g = \mathcal{N}(0, \sigma^2/K)$, $\mathcal{E}_{\mathrm{A}}$ is the statistical expectation w.r.t. the distribution $\mathcal{N}(A, \sigma^2/K)$, We use Monte Carlo simulation to estimate $\mathcal{E}_{\mathrm{A}}\left\{I_{\{\bar{w}>A\}} \exp\left[\left(-2A\bar{w} + A^2\right) / \left(2\sigma^2/K\right)\right]\right\}$, the results are shown in Table 2, the number of experiments in Monte Carlo simulation is 1000, the 95% confidence intervals are also included in the table. The Monte Carlo simulation results fit well with the theoretical result (7).

5 Conclusion and Future Work

For detection of speech polarity, the speech residual signal can be modeled as impulse train plus WGN, the optimal detector outperformes the original polarity detection algorithm by tens of order of magnitude in term of detection error. This result is validated by Monte Carlo simulation. It should be noted that in the above analysis, we have assumed that the parameters of the impulse train P, A and the AR model parameters are all assumed known, in practice, these parameters must be estimated from the speech signal, the estimation error will degrade the detector performance. The performance loss using estimated parameters is under investigation and will be reported in a future paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 60272074 and the Spaceflight Innovation Foundation of China under grant [2002]210-6.

References

[1] S. Sakaguchi, T. Arai and Y. Murahara. The Effect of Polarity Inversion of Speech on Human Perception and Data Hiding as an Application. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, June 5-9, (2000) **2** 917-920

[2] Yardyimci, Y., Cetin, A. E., and Ansari, R.: Data Hiding in Speech Using Phase Coding. Eurospeech 97 **3** (1997) 1679-1682

[3] Ciloglu, T. and Karaaslan, S. Utku: An Improved All-Pass Watermarking Scheme for Speech and Audio. International Conference on Multimedia and Expo. July 30- Aug. 2 (2000) **2** 1017-1020

[4] Cheng, Q., Sorensen, J., "Spread Spectrum Signaling For Speech Watermarking", IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA,pp. 1337-1340.

[5] J. R. Deller, Jr. , J. G. Proakis, J. H. L. Hansen, Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, USA, 1993

[6] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory. Prentice-Hall, New Jersey, 1998

[7] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification (2nd Edition), Wiley-Interscience, USA, 2001

[8] V. Barnett, T. Lewis. Outliers in Statistical Data, John Wiley and Sons, New York, 1994.

[9] H. J. Kim, T. Kim, In-Kwon Yeo. A Robust Audio Watermarking Scheme. IEEE ISCAS, Canada, May 2004.

[10] Peter J. Smith, M. Shafi, Hongsheng Gao. Quick Simulation: A Review of Importance Sampling Techniques in Communications Systems. IEEE Journal on Selected Areas in Communications, May 1997, Vol.15, No. 4, 597-613

[11] R. L. Mitchell. Importance Sampling Applied to Simulation of False Alarm Statistics. IEEE Trans. on Aerosp. Elect. Syst., Jan, 1981, Vol. AES-17, 15-24