INTERSPEECH 2009 BRIGHTON

Watermark Recovery From Speech Using Inverse Filtering And Sign Correlation

Robert Morris[1], *Ralph Johnson*[1], *Vladimir Goncharoff*[2], *and Joseph DiVita*[1]

[1]SPAWAR Systems Center Pacific, 53560 Hull St., San Diego, CA 92152
[2]ECE Dept., University of Illinois at Chicago, Chicago, IL 60607

rob.morris@navy.mil, ralph.johnson@navy.mil, volodia@uic.edu, joseph.divita@navy.mil

10.21437/Interspeech.2009-409

Abstract

This paper presents an improved method for asynchronous embedding and recovery of sub-audible watermarks in speech signals. The watermark, a sequence of DTMF tones, was added to speech without knowledge of its time-varying characteristics. Watermark recovery began by implementing a synchronized zero-phase inverse filtering operation to decorrelate the speech during its voiced segments. The final step was to apply the sign correlation technique, which resulted in performance advantages over linear correlation detection. Our simulations include the effects of finite word length in the correlator.

Index Terms: Speech Watermarking, Hidden Tones, Speech Steganography, Speech Data Hiding, Sign Cross Correlation, Inverse Filtering

1. Introduction

Watermarking speech signals may be done for the purposes of copyright protection, time-stamping, tamper detection, or embedding any sort of side information. Typically it is required that the presence of such watermarks not affect the perceived quality of the speech signal. Hiding the watermark is made easier by taking advantage of time-varying speech characteristics. For example, as done with quantization noise in perceptual audio coders, the watermark level may be made to vary with short-time speech energy in order to keep the watermark below a perceptual threshold of detection. Such approaches have included hiding watermark energy under speech formants [1] and inserting short tones at frame-by-frame computed levels [2].

In a previous paper we presented a method to detect a watermark, in the form of a sequence of DTMF tones, after it had been added to speech at sub-audible levels [3]. There, no knowledge of the speech signal at time of watermarking was assumed - the watermark was added to the speech asynchronously. Also, the watermark was detected without knowledge of the original un-watermarked speech (blind detection). Under these constraints the speech signal may be thought of as a noise that interferes with watermark detection, and any detection method must employ time averaging to achieve processing gain. Further, finite word lengths in the correlator add a noise of their own that measurably affects its performance. We had investigated three different nonlinearities in an attempt to improve watermark detection capability using simple bit-level manipulations of signed integer codewords. The most promising of these methods was to replace all bits in the speech-plus-watermark codeword with copies of the sign bit. Here we describe a technique to improve processing gain in our DTMF tone watermarking system, using zero-phase inverse filtering prior to the sign bit manipulation.

2. Background

2.1. Watermark Embedding

Our speech watermarking scheme adds a low-level watermark to a speech signal directly in the time domain:

$$y(n) = s(n) + \lambda w(n), \tag{1}$$

where $s(n)$ is the zero-mean speech signal, $w(n)$ is the zero-mean watermark signal, and scaling factor λ is selected to adjust the input signal-to-noise ratio. We define SNR_{in} from the point of view that $w(n)$ is the signal and $s(n)$ is the interfering "noise": $\lambda^2\sigma_w^2/\sigma_s^2$.

In a linear correlation watermark detector, the presence of a specific watermark is detected by cross-correlating $w(n)$ and $y(n)$:

$$\begin{aligned} R_{wy} &= E[w(n)y(n)] \\ &= E[w(n)\,(s(n)+\lambda w(n))] = \begin{cases} \lambda\sigma_w^2 & w(n) \text{ present} \\ 0 & w(n) \text{ absent} \end{cases} \end{aligned}$$

In practice the value of R_{wy} may be estimated by averaging N available sample points to obtain the sample mean of $z(n) = w(n)y(n)$:

$$M_N(z) = \frac{1}{N}\sum_{n=1}^{N} z(n) = R_z + \varepsilon_z(N),$$

where $\varepsilon_z(N)$ is the error due to averaging only N samples. As $N \to \infty$ averaging noise ε_z approaches zero. It may be shown that the variance of $\varepsilon_z(N)$ is equal to σ_z^2/N, leading to an expression for signal-to-noise ratio at the output of the linear correlation detector:

$$SNR_{out} = \frac{R_z^2}{\sigma_z^2/N} = N\frac{R_{wy}^2}{\sigma_{wy}^2}$$

Averaging noise imposes practical limits on watermark detection: SNR_{in} cannot be decreased below a certain level, otherwise the signal component (R_{wy}) at the correlator output will become obscured by the noise component (ε_z). Higher values of processing gain $G = SNR_{out}\,/\,SNR_{in}$ make it possible to reduce SNR_{in} (better hiding the watermark) while still being able to recover the watermark.

2.2. Sign Correlation

We had previously reported the effect of various instantaneous nonlinearites, applied to $y(n)$ prior to correlation, on minimum

SNR_{in} for DTMF watermark detection [3]. One of these nonlinear operations was the sign operator: a significant increase in processing gain resulted from replacing y with $\hat{y} = sign(y)$ for speech watermark detection. This method is referred to as *sign correlation*, shown to approximate the log-likelihood ratio measure when dealing with uncorrelated Laplacian-distributed signals with additive watermark [4]:

$$\frac{1}{N}\sum_{n=1}^{N} w(n)\, sign\{s(n)+\lambda w(n)\} = R_{w\hat{y}} + \varepsilon_{w\hat{y}}(N)$$

The sign correlator output signal component $R_{w\hat{y}}$ is bounded to the range $[0, E[|w|]]$, which is advantageous for selecting a detection threshold value when the exact value of λ (and hence SNR_{in}) is not known. More importantly, in our experiments with additive DTMF watermarks in speech, sign correlation raised processing gain by approximately 26 dB over that given by linear correlation.

To analyze what happens with sign correlation, assume speech sequence $s(n)$ and watermark sequence $w(n)$ are defined only over $1 \le n \le N$, and the watermarked speech signal is $y(n) = s(n) + \lambda w(n)$. The sign correlation between $w(n)$ and $y(n)$ is $E[w(n)\, sign\{y(n)\}]$. We may estimate the expected value using an N-point sample mean. Let $\hat{z}(n) = w(n)\, sign\{y(n)\}$:

$$\begin{aligned} E[\hat{z}(n)] \approx M_N(\hat{z}) &= \frac{1}{N}\sum_{n=1}^{N}\hat{z}(n) \\ &= \frac{1}{N}\sum_{n=1}^{N} w(n)\, sign\{y(n)\} \\ &= \frac{1}{N}\sum_{n=1}^{N} w(n)\, sign\{s(n)+\lambda w(n)\} \end{aligned}$$

Express $M_N(\hat{z})$ as a sum of two terms

$$M_N(\hat{z}) = A_1 + A_2,$$

where:

1. $A_1 = \frac{1}{N}\sum_{n\in T_1}\hat{z}(n)$ (T_1 is the set of all time index values n for which $s(n)$ and $w(n)$ are the same sign), and
2. $A_2 = \frac{1}{N}\sum_{n\in T_2}\hat{z}(n)$ (T_2 is the set of all time index values n for which $s(n)$ and $w(n)$ are of opposite sign).

Define $N_1 = size\{T_1\}$ and $N_2 = size\{T_2\}$. Note that $N = N_1 + N_2$ and $N_1 \approx N_2 \approx N/2$, because the speech and watermark signals are statistically independent and each is assumed to be positive or negative with equal probability.

First consider A_1:

$$\begin{aligned} A_1 &= \frac{1}{N}\sum_{n\in T_1} w(n)\, sign\{s(n)+\lambda w(n)\} \\ &= \frac{1}{N}\sum_{n\in T_1} w(n)\, sign\{w(n)\} \\ &= \frac{1}{N}\sum_{n\in T_1} |w(n)| = \frac{N_1}{N} M_{N_1}(|w|). \end{aligned}$$

Thus $A_1 \approx M_{N_1}(|w|)/2$ and is independent of parameter λ. The variance of A_1, nonzero as a result of averaging only N_1 samples, is $Var\left[\frac{N_1}{N}M_{N_1}(|w|)\right] = \frac{N_1^2}{N^2}\left(\frac{\sigma^2_{|w|}}{N_1}\right) \approx \frac{\sigma^2_{|w|}}{2N}$.

Next consider the term A_2, which unlike A_1 is a function of λ:

$$\begin{aligned} A_2(\lambda) &= \frac{1}{N}\sum_{n\in T_2} w(n)\, sign\{s(n)+\lambda w(n)\} \\ &= \frac{N_2}{N} M_{N_2}\left(w(n)\, sign\{s(n)+\lambda w(n)\}\right) \end{aligned}$$

As λ varies from 0 to ∞, A_2 varies from $\frac{N_2}{N}M_{N_2}(-|w|)$ to $\frac{N_2}{N}M_{N_2}(|w|)$. In a speech watermarking application it is typically the case that $\lambda^2\sigma_w^2 \ll \sigma_s^2$ ($SNR_{in} \ll 1$), for which $A_2 \approx \frac{N_2}{N}M_{N_2}(-|w|) \approx -M_{N_2}(|w|)/2$. In such case the variance of A_2, nonzero as a result of averaging only N_2 samples, is $\approx \frac{\sigma^2_{|w|}}{2N}$.

Assuming statistical independence between the data points used to calculate estimates A_1 and A_2, the corresponding estimation error noises may be added together to obtain the total output noise power $\sigma^2_{|w|}/N$.

In the case of sign correlation watermark detection, the best possible case is when all samples of the speech plus watermark signal have the same sign as the watermark. Then the signal component of cross-correlation result will equal $E[|w|]$, or $\mu_{|w|}$. Thus the maximum possible output signal power for a sign correlation detector is $\mu^2_{|w|}$, giving

$$SNR_{out}(max) = \frac{\mu^2_{|w|}}{\sigma^2_{|w|}/N}$$

For example, for sinusoidal watermark $SNR_{out}(max) = 4.28N$. Watermarks comprised of DTMF tones have 3.5 dB lower $SNR_{out}(max) = 1.92N$. Thus $SNR_{out}(max)$ of the sign correlator is fixed by the choice of watermark and the number of available samples for averaging.

Our previous work had simulated the functionality of $sign(y)$, when y is represented using signed integer code, by replacing all bits in the codeword with copies of the sign bit (REM function, parameter $k = 0$) [3]. Instead of $sign(y)$ this operation yields $sign(y+\varepsilon) \times const$, where ε is a positive offset corresponding to half the weight of the least significant bit in the code for y. As a result of offset ε being added to y, a threshold level is established below which a finite-amplitude-span watermark cannot be detected.

2.3. Zero-phase inverse filtering

Here we report that an additional ~ 5 dB processing gain improvement is achieved, in the case that y is stored using 16-bit signed integer codewords, by decorrelating $y = s + \lambda w$ with an adaptive FIR zero-phase-shift inverse filtering operation prior to sign correlation.

Consider the time-invariant linear filtering operation

$$g_1(n) = a(n) * y(n) = \sum_{k=0}^{p} a(k)y(n-k),$$

where the impulse response $a(n)$ is finite-length and causal. This filter introduces a phase shift that may be undone by passing a time-reversed copy of $g_1(n)$ through the same filter, then time-reversing the result:

$$\begin{aligned} g_2(n) &= a(n) * g_1(-n) \\ &= a(n) * a(-n) * y(-n); \end{aligned}$$

$$g_2(-n) = a(-n) * a(n) * y(n) = \sum_{k=-p}^{p} h(k)y(n-k),$$

where the finite-length impulse response of the overall filter is $h(k) = a(-k) * a(k)$. The overall filter's frequency response is $H(e^{j\omega}) = \left|A(e^{j\omega})\right|^2$, having magnitude response that is the original magnitude response squared, and a phase shift of zero. Note that $h(k) = h(-k)$, so that this filter is not causal.

In our work we obtained the original filter parameters $a(n)$ in a time-dependent manner as the optimal *AR(12)* model coefficients matching $\sqrt{|Y_n(e^{j\omega})|}$, where $Y_n(e^{j\omega})$ is a short-time spectrum derived from a 20 msec raised-cosine-windowed section of y in the vicinity of time index n. Thus the frequency response of our 24^{th}-order filter, $H_n(e^{j\omega}) \approx 1/|Y_n(e^{j\omega})|$, is a zero-phase-shift inverse filter adapting to the short-time characteristics of $y(n)$ and whitening it by removing the energy of linearly predictable components. The zero phase shift characteristic of the inverse filter was found to be necessary to minimize the linear distortion imparted on the watermark component of $y(n)$; otherwise the ability to detect the watermark suffered in our experiments. By inspecting the results of this filtering, it was obvious that the high-energy voiced regions of speech were most affected. We believe this improvement in watermark-to-speech power ratio in voiced regions explains the improved overall processing gain of our sign correlator watermark detection system.

3. Watermark Recovery

3.1. Watermark Generation

In the experiments which follow, the watermark signal w was derived from a sequence of P DTMF tones

$$\theta_P = \left[\bar{d}_1, \ldots, \bar{d}_P\right] \tag{2}$$

where each DTMF tone $\bar{d}_i \in \Re_{16}^{1\times K}$ had a duration of 100 milliseconds (i.e. $K = f_s/10$, for a sample rate of f_s). Since there are 16 available DTMF tones, a total of 16^P unique DTMF sequences could be generated. The watermark

$$w = [\theta_P^{(1)}, \ldots, \theta_P^{(q)}]^T \tag{3}$$

was then constructed by repeating θ_P until the length of the watermark (qKP) was equal to the number of samples in $\hat{s}$. Note that the original speech signal, s, is now truncated to $\hat{s}$; a segment whose length is a multiple of KP to match the DTMF sequence.

3.2. Correlation Detection Score

The Correlation Detection Score (CDS) along with a threshold, α, provides a test to determine whether the watermark is present in the speech signal. The method returns a continuous range of values $[0, 5]$ where a higher value demonstrates a higher level of detection confidence.

Assuming that w and y are independent and that either the expected value of the watermark or the speech is zero, the true cross-correlation

$$\begin{aligned} R_{wy}(m) &= E\left[w(n+m)y(n)\right] \\ &= E\left[w(n+m)(\hat{s}(n) + \lambda w(n))\right] \\ &= E\left[w(n+m)\right] E\left[\hat{s}(n)\right] + \lambda E\left[w(n+m)w(n)\right] \\ &= \lambda E\left[w(n+m)w(n)\right] = \lambda R_{ww}(m) \end{aligned}$$

is equal to a constant times the autocorrelation of the watermark signal. Hence, using the N-point sample mean

$$\begin{aligned} R_{wy}(m) \approx M_N(w(n+m)y(n)) &= \tilde{R}_{wy}(m) \\ &= \lambda R_{ww}(m) + \varepsilon_{ww}(N), \end{aligned}$$

if the averaging error $\varepsilon_{ww}(N)$ is small, it is expected that $\tilde{R}_{wy}$ will be close to the scaled autocorrelation of the watermark w. Thus, the correlation detection score is derived in order to detemine how well the estimate $\tilde{R}_{wy}$ matches the scaled autocorrelation λR_{ww}, which is known a priori.

Since the reference correlation R_{ww} is an even function, the information in the left and right halves is equivalent. Therefore only the coefficients in the left half

$$c_{wy}(m) = \tilde{R}_{wy}(m - N + KP/2), m = 1, \ldots, N$$

were considered in the scoring function. Note that the coefficients are shifted to the right by half of the length of θ_P so that windowing can be centered around each correlation peak. Finally, the correlation is squared and normalized to produce the correlation sequence

$$\tilde{c}_{wy}(m) = \frac{c_{wy}(m)^2}{\max_{1\le k\le N}(c_{wy}(k)^2)}, m = 1, \ldots, N$$

which becomes independent of λ because of the normalization.

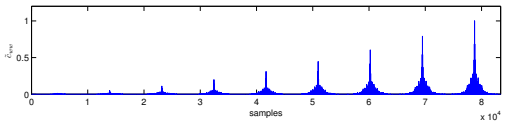
Define $i_1, \ldots, i_q$ to be the q peak indices of the autocorrelation sequence $\tilde{c}_{ww}(m), m = 1, \ldots, N$, corresponding to when the individual watermarks ($\theta_P^{(i)}$) align with each other. First the raw score

$$\Psi_j = \begin{cases} 1 & \text{if } i_j = \arg\max_{[i_j - KP/4 \le m \le i_j + KP/4]} \tilde{c}_{wy}(m) \\ 0 & \text{otherwise} \end{cases}$$

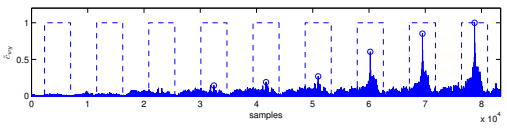
was determined for each of the q autocorrelation peaks. The correlation detection score is then calculated as

$$S_{wy} = \beta \sum_{j=1}^{q} \tilde{c}_{ww}(i_j)\Psi_j$$

where the amplitude of the peaks $\tilde{c}_{ww}(i_j)$ are used as weighting factors and $\beta = \frac{5}{\sum_{j=1}^{q} \tilde{c}_{ww}(i_j)}$ scales the score between 0 and 5. Since the peak amplitudes follow a triangular shape (see Figure 1a) the weights were designed to reward the higher valued peaks which are less likely to be dominated by adjacent noise.



(a) $\tilde{c}_{ww}$



(b) $\tilde{c}_{wy}$ with 6 matching peak locations.

Figure 1: Determining the Correlation Detection Score of speech with a -30 dB watermark.

A cross-correlation sequence $\tilde{c}_{wy}$ between the watermark and y, illustrated in Figure 1b, is detected by comparing the constrained peak locations with the corresponding peak locations of the autocorrelation sequence $\tilde{c}_{ww}$ shown in Figure 1a. The broken lines indicate the constraint placed on each peak and the circles at the peaks of $\tilde{c}_{wy}$ indicate when the highest peak within each window matches the corresponding peak location of $\tilde{c}_{ww}$. In this case, only six peaks matched giving a correlation detection score $S_{wy} = 4.7544$.

4. Experimental Results

The following sections demonstrate performance of the *sign cross correlation* enhancement method using 8 kHz clean speech and a watermark created from a sequence of DTMF tones described earlier in Section 3.1. For each experiment, a 1-sec DTMF sequence was created (see Eq. 2) using the tones from the ten digit sequence "123456789A", and added to each speech segment by repetition via the construction in Eq. (3).

4.1. *Sign Cross Correlation* Method with Speech

To demonstrate performance of the sign cross correlation method, twenty male speakers from the TIMIT database were selected at random. Utterances from each speaker were concatenated after discarding all but 0.1 seconds of header and trailer silence. In previous work [3], this utterance beginning and ending silence was retained. However, in this experiment, it was desired to determine whether the watermark recovery was enabled by these extra silence periods. Removing the entire silence periods, on the other hand, would produce unnatural flow between sentences. So 0.1 seconds was arbitrarily selected placing 0.2 seconds of silence between each TIMIT utterence in the concatenated sequence.

Since the DTMF nature of the watermarks is telephone bandwidth, the concatenated sequences were downsampled from 16 kHz to 8 kHz. The total speech duration per speaker was used to generate progressively longer speech segments $\hat{s}_2^i, \hat{s}_4^i, \ldots, \hat{s}_{24}^i$ where the subscript indicates the duration in seconds and i is the speaker ID. The 1-sec DTMF watermark sequence was added to each $\hat{s}_j^i$ by repetition.

The lowest detection level (using $\alpha = 2$) was calculated for each speaker segment $\hat{s}_j^i, j = 2, 4, \ldots, 24; i = 1, \ldots, 20$. The mean, over the speakers, is plotted in Figure 2 as the durations

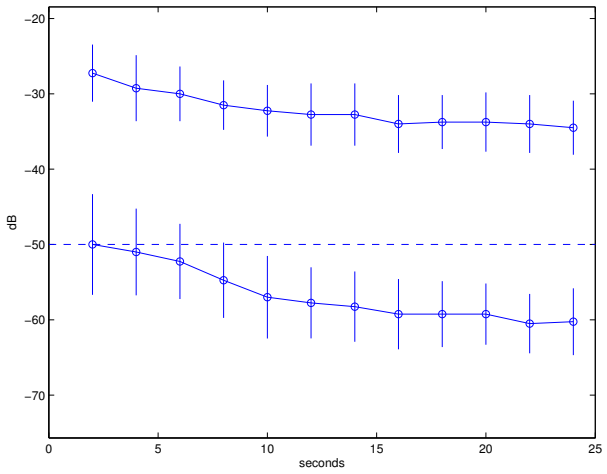


Figure 2: Evaluation of *sign cross correlation* method on DTMF watermark over 20 TIMIT speakers using varying durations.

are increased. The upper line in Figure 2 shows the lowest detection level without enhancement, the broken line approximates the human detection threshold, as determined with informal listening tests, and the lower line shows the detection level when the *sign* of the speech plus watermark is used to enhance the correlation. An average of 26 dB improvement is seen after the enhancement, which is comparable to using the higher sampling rate and full headers as was done in [3]. The vertical lines at each data point indicate the range of plus or minus σ across the 20 TIMIT speakers.

4.2. Adding Inverse Filtering with Zero Phase Shift

Using the same experiment setup as Section 4.1, an inverse filter with zero phase shift was applied to the speech plus watermark prior to sign cross correlation. The lower line in Figure 3 displays an additional 5 to 6 dB mean improvement in the lowest detection levels compared to the sign cross-correlation line that is shown without whiskers for clarity. The human perception and unenhanced lines are displayed for reference.

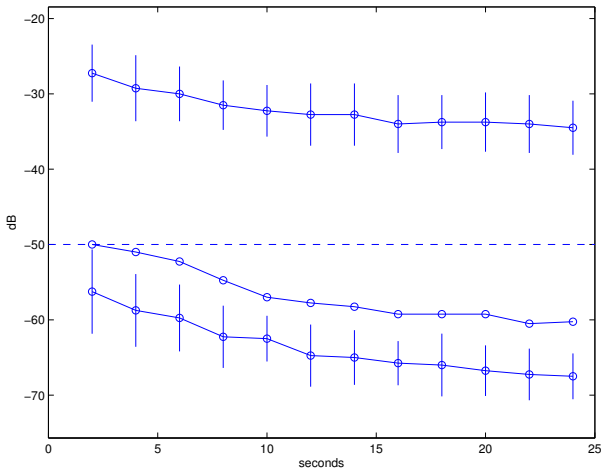


Figure 3: Evaluation of *sign cross correlation* method after applying the zero phase shift inverse filter on DTMF watermark over 20 TIMIT speakers using varying durations.

5. Conclusion

Sign correlation has been shown to have approximately 26 dB greater processing gain than does linear correlation in detecting low-energy DTMF watermarks that have been asynchronously added to speech. An additional 5 to 6 dB improvement was achieved using zero-phase-shift inverse filtering prior to sign correlation. Our experiments took into account the effects of finite word length in the sign correlator.

6. Acknowledgements

This work was supported by the Office of Naval Research through the In-House Laboratory Independent Research program at SPAWAR Systems Center Pacific.

7. References

[1] Qiang Cheng and Jeffrey Sorenson, "Spread spectrum signaling for speech watermarking," *Proc. IEEE ICASSP*, pp. 1337–1340, 2001.

[2] Kaliappan Gopalan and Stanley Wenndt, "Audio steganography for covert data transmission by imperceptible tone insertion," *Proceedings Communications Systems and Applications, IEEE*, vol. 4, pp. 1647–1653, 2004.

[3] Robert Morris, Ralph Johnson, Vladimir Goncharoff, and Joseph DiVita, "Recovering asynchronous watermark tones from speech," *Proc. IEEE ICASSP*, 2009.

[4] Xiaochen Bo, Lincheng Shen, and Wensen Chang, "Sign correlation detector for blind image watermarking in the dct domain," *Proc. IEEE PRC on Multimedia*, vol. 2, pp. 780–787, 2001.