RECOVERING ASYNCHRONOUS WATERMARK TONES FROM SPEECH

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
rob.morris@navy.mil, ralph.johnson@navy.mil, joseph.divita@navy.mil
[2]Dept. of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, Ill 60607
volodia@uic.edu

ABSTRACT

A new, low complexity method facilitates low burden embedding and recovery of tonal watermarks in speech. A watermark composed of a periodically extended sequence of sub-audible DTMF tones is added to speech asynchronously, without regard to momentary speech characteristics. It is detected through a combination of a bit manipulation enhancement and a data-directed correlation, ideal for simple hardware implementations. Three methods of bit manipulation enhancement were auditioned and the best selected for further investigation. It showed an average 26 dB processing gain vs. correlation alone, sufficient to detect the asynchronous sub-audible tones by a comfortable margin.

Index Terms— Speech Watermarking, Hidden Tones, Speech Steganography, Speech Data Hiding

1. BACKGROUND

Imperceptibly embedded data can be used to stamp speech with a watermark. In many applications the watermark must be transparent to the listener of the speech content, and should not rob any power from the signal or affect its content by noticeably changing the speech power level or its intelligibility. Additionally, it would be ideal to minimize any delay, processing load, or system modification burden at the point of watermark generation and insertion. It would also be desirable to have a low complexity recovery method.

Prior researchers' approaches have included directly replacing the lower bits in PCM samples [1], replacing the unvoiced CELP residual [2], impressing coded phase changes onto the analog waveform, hiding spread spectrum under formants [3], and inserting short tones at frame by frame computed levels [4].

Many of those approaches tried to minimize the difficulty in watermark recovery by maximizing the watermark power. That was done by inserting data piecemeal at higher power levels, skirting the threshold of hearing and the limits of perceptual masking. These methods attempt to mask data by inserting it only into certain strongly voiced speech segments, or by inserting it all throughout speech, but at custom power ratios calculated for each short segment. These approaches require processing buffer delays that preclude real-time, instantaneous encoding. They also require considerable processing load, both at the insertion stage and at the recovery.

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

2. INTRODUCTION

The proposed new method allows instantaneous encoding through a simple mixing of DTMF tones. It adds the tones asynchronously, without any knowledge of the momentary speech details, or of any piecemeal speech/data power relationships.

Human perception is quite sensitive to tones, particularly in very clean speech, so they must be inserted at a very low level, making recovery extremely difficult. Informal listening found the tones inaudible at a roughly -50 dB power level.

The new recovery method has two components: preprocessing by bit manipulations, and a data-directed correlation. This paper compares the detection by correlation alone to that after enhancement by a low complexity method.

An extra benefit of this scheme is that the calculation and analysis load is borne essentially by the detection/recovery process, with minimal burden at the encoding end. That also means that minimal technical equipment changes are needed to add watermarks, and that any significant changes are required for only those interested in detecting or decoding the watermark.

2.1. Watermark Embedding

Assume that a watermark signal is scaled and added to a truncated speech signal

$$y = \hat{s} + \lambda w \in I_{16}^{N\times 1} \quad (1)$$

where $\hat{s} \in I_{16}^{N\times 1}$ is the speech signal represented as a 16-bit signed integer code, $\lambda \in \Re$ is a scaling factor, and $w \in I_{16}^{N\times 1}$

is the watermark. In general, λ is independent of $\hat{s}$. When the speech signal is available, the value for λ may be calculated

$$\lambda = \left[\frac{\sum_{n=1}^{N} (\hat{s}_n)^2}{\sum_{n=1}^{N} (w_n)^2} 10^{r/10} \right]^{1/2}$$

where $\hat{s}_n$ and w_n are the components of the speech and watermark signal, and r is a desired watermark to speech power ratio in dB. If the speech signal is not available, the value of λ can be determined by an arbitrary estimate of the power of an average speech signal.

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 $\bar{d}_i \in I_{16}^{1 \times K}$ 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, was truncated to $\hat{s}$; a segment whose length is a multiple of KP to match the DTMF sequence.

2.2. Correlation Analysis

The true cross-correlation sequence between the watermark and the speech is

$$R_{wy}(m) = E\left[w_{n+m} y_n \right] \tag{4}$$

where w_n and y_n are stationary random processes representing the watermark and speech plus watermark respectively, $-\infty < n < \infty$, and $E[\cdot]$ is the expectation operator. Assuming that w and y are independent and that either the expected value of the watermark or the speech is zero, using Eq. (1) the cross-correlation

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

is equal to a constant times the autocorrelation of the watermark signal.

3. ANALYSIS OF RECOVERY METHODS

3.1. Preprocessing by Bit Manipulation

In practical application, a sample mean is used to estimate the expectation operator in Eq. (4):

$$E\left[w_{n+m} y_n \right] \approx M_N(w_{n+m} y_n) = \lambda R_{ww}(m) + e, \tag{5}$$

where e is the estimation error that results from substituting $E\left[w_{n+m} y_n \right]$ with $M_N(w_{n+m} y_n) = \frac{1}{N} \sum_{n=1}^{N} w_{n+m} y_n$. Since

$$\begin{aligned} M_N(w_{n+m} y_n) &= M_N(w_{n+m}(\hat{s}_n + \lambda w_n)) \\ &= M_N(w_{n+m} \hat{s}_n) + \lambda M_N(w_{n+m} w_n), \end{aligned}$$

we see that $e = e_1 + e_2 : e_1 = E\left[w_{n+m} \hat{s}_n \right] - M_N(w_{n+m} \hat{s}_n)$ and $e_2 = \lambda(E\left[w_{n+m} w_n \right] - M_N(w_{n+m} w_n))$. The law of large numbers states that $\sigma_{e_1}^2 = \sigma_{w_{n+m}\hat{s}_n}^2 / N$ and $\sigma_{e_2}^2 = \lambda^2 \sigma_{w_{n+m} w_n}^2 / N$, and since the watermark signal λw_n is intentionally many decibels below speech $\hat{s}_n$ in power level we may presume that $\sigma_e^2 \approx \sigma_{e_1}^2$. Therefore once the waveforms and parameters $\{w_n, \hat{s}_n, \lambda, N\}$ are selected one may attempt to reduce σ_e^2 by reducing the variance of $w_{n+m} \hat{s}_n$ through some kind of nonlinear processing prior to correlation.

Our work has been to apply three different instantaneous nonlinearities to the watermarked speech, $y_n = \hat{s}_n + \lambda w_n$, in order to improve the resulting estimate of the autocorrelation function $R_{ww}(m)$. To ensure computational efficiency, each of the three nonlinear preprocessing methods are shown below to have simple implementations using bit-level manipulations on signed integer (also known as 2's-complement) binary codes.

The first method that we have investigated for improving watermark in speech recovery we have called the REM method. It gets its name from the remainder function that defines it as

$$REM(y_n, k) = rem(y_n + \frac{1}{2}, 2^k) - \frac{1}{2}.$$

With signed integer codes, the REM method is implemented as follows: retain the k least-significant bits without any change, and replace all other bits with copies of the sign bit. The second method is an amplitude limiting process

$$AL(y_n, k) = sign(y_n) \cdot min(|y_n|, 2^k).$$

With signed integer codes the AL method is implemented as follows: if all except the k right-most bits are the same in value, then make no change. Otherwise clear the k right-most bits, set the bit to their left, and replace all other bits with copies of the sign bit. Finally, the third of our processing methods is the $SIGN$ method:

$$SIGN(y_n) = [y_n \geq 0] - 1,$$

where the test for $y_n \geq 0$ returns 1 if true and 0 if false. When applied on signed integer codes, all bits are replaced with copies of the sign bit. It should be noted that both the $SIGN$ and REM methods introduce a d.c. bias that may be subtracted if desired.

The following figure shows the relative processing gain resulting from all three methods on a sum of a zero-mean, Gaussian random watermark when scaled to be 50 dB below a

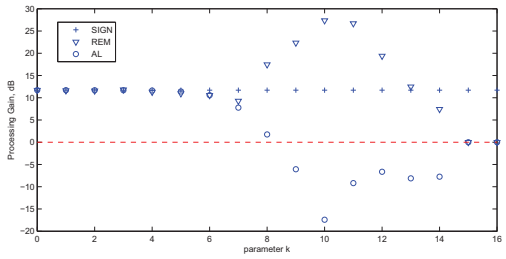


Fig. 1. Processing gain while comparing $SIGN$, REM, and AL methods.

100Hz-tone model for speech ($N = 10^6$). We have found the nonlinear processing effectiveness in improving output SNR to be very much signal-dependent. The plot above shows an experimental result where the REM method with parameter $k = 10$ achieved in excess of 25 dB processing gain compared to cross-correlation without any nonlinear preprocessing.

3.2. Data-Directed Watermark Detection

The data-directed correlation detection method along with a threshold, α, provides a test to determine whether the watermark is present in the speech signal. Using a modified correlation, the method returns a continuous range of values between 0 and 5 where the higher value demonstrates a higher level of detection confidence.

The Correlation Detection Score (CDS) is a measure of the quality of the cross-correlation between w and y as compared to the autocorrelation of the watermark w. When the error e is small (see Eq. 5), it is expected that R_{wy} will be close to the scaled autocorrelation of the watermark. Therefore, an objective measure was derived which determines how well 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

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

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.

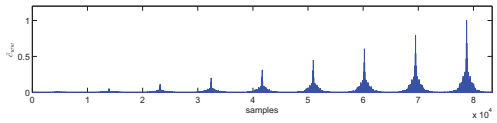
Define $i_1, \ldots, i_q$ to be the q peak indices of the autocorrelation sequence $\tilde{c}_{ww}(m), m = 1, \ldots, N$, corresponding to when the individual watermarks ($\theta_P^{(i)}$) align with each other. First the raw score

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

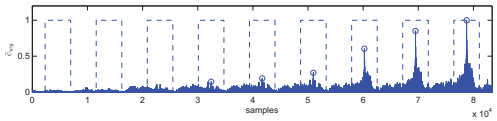
was determined for each of the q autocorrelation peaks. The correlation detection score is then calculated as

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

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 2a) 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.

Fig. 2. Determining the Correlation Detection Score of speech with a -30 dB watermark.

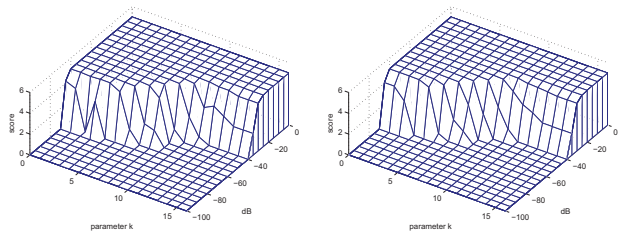
A cross-correlation sequence $\tilde{c}_{wy}$ between the watermark and y, illustrated in Figure 2b, is detected by comparing the constrained peak locations with the corresponding peak locations of the autocorrelation sequence $\tilde{c}_{ww}$ shown in Figure 2a. The broken lines indicate the constraint 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 REM, AL, and $SIGN$ enhancement methods using 16 kHz clean speech and a watermark created from a sequence of DTMF tones described earlier in Section 2.1. 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. REM, AL, and $SIGN$ Methods with Speech

A male speaker from the TIMIT database was selected at random and the speech from his ten utterances was concatenated up to a total duration of 30-sec. After the DTMF watermark was added at a varying signal to noise ratio, the CDS was determined as the value of parameter k was modified. The results for the REM and AL method appear in Figure 3a and 3b below. The performance of both is similar: decreasing k enables one to detect a weaker watermark signal. The $SIGN$ method exceeds or equals the performance of the other two methods for every value of k (30 dB gain compared with no enhancement). Note, when $k = 0$: $REM(y_n, k) = SIGN(y_n)$, and $AL(y_n, k)$ differs from the other two only for $y_n = 0$ (when the three nonlinear functions are normalized to have the same amplitude range). Because of this, the $SIGN$ method was chosen for further investigation.



(a) REM (b) AL

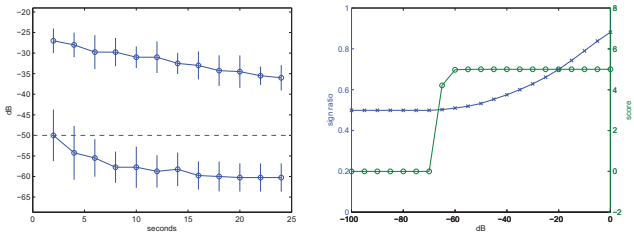
Fig. 3. Correlation Detection Score (CDS) as watermark dB level and number of bits are varied.

4.2. $SIGN$ Method with Multiple Speakers

To demonstrate the improvement over a wider range of speech samples, performance was evaluated for 20 randomly selected male TIMIT speakers. Utterances from each speaker were concatenated and the total speech duration per speaker was used to generate progressively longer speech segments $\hat{s}^i_2, \hat{s}^i_4, \ldots, \hat{s}^i_{24}$ where the subscript indicates the duration in seconds and i is the speaker ID. The 1-sec DTMF sequence was added to each $\hat{s}^i_j$ by repetition.

The lowest detection level (using $\alpha = 2$) was calculated for each speaker segment $\hat{s}^i_j, j = 2, 4, \ldots, 24; i = 1, \ldots, 20$. The mean, over the speakers, is plotted in Figure 4a as the durations are increased. The upper line in Figure 4a shows the lowest detection level without enhancement, the broken line approximates the human detection threshold, and the lower line shows an average of 26 dB improvement after enhancement. The vertical lines at each data point indicate the range of plus or minus σ among the 20 TIMIT speakers.

Also seen in Figure 4a is that as the speech segment duration doubles, the SNR detection level gains approximately the expected 3 dB. However, the last 5 samples of the enhanced plot line indicate that an asymptote is reached at near -60 dB.



(a) Twenty speakers. (b) Single speaker.

Fig. 4. Evaluation of $SIGN$ method.

This can be explained because the $SIGN$ method requires that the ratio $\gamma = \frac{1}{N}\sum_{n=1}^{N}[SIGN(\hat{y}_n) == SIGN(\lambda w_n)]$ must not represent random chance. Varying the watermark dB level on a single 30-sec TIMIT speech file (Figure 4b), it can be seen that as the signal to noise ratio is reduced γ approaches 0.5. Also note that the corresponding score drops to zero near the input SNR level where γ reaches the asymptote.

5. CONCLUSION

An imperceptible tonal watermark can be embedded in speech asynchronously and detected using unique combinations of bit manipulation enhancement along with a data-directed correlation. This watermarking method meets the desired criteria: transparent to listeners, minimal burden at insertion, no significant change in the speech communication power, and low complexity recovery. It is ideal for implementation in simple hardware. Under certain circumstances, REM produced better performance when compared to the other methods, however, in the speech experiments performed, REM did not exceed the $SIGN$ method.

6. REFERENCES

[1] Chung-Ping Wu and C. C. Jay Kuo, "Fragile speech watermarking for content integrity verification," *Proc. IEEE ICASSP*, vol. 2, pp. 436–439, 2002.

[2] Chia-Hsiung Liu and Oscal T. C. Chen, "A fragile watermarking scheme with recovering speech contents," *The 47th IEEE International Midwest Symposium on Circuits and Systems*, vol. 2, pp. 165–168, 2002.

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

[4] 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.