SPREAD SPECTRUM SIGNALING FOR SPEECH WATERMARKING

Qiang Cheng

University of Illinois
Urbana Champaign, IL

Jeffrey Sorensen

IBM T. J. Watson Research Center
Yorktown Heights, NY
sorenj@us.ibm.com

ABSTRACT

The technique of embedding a digital signal into an audio recording or image using techniques that render the signal imperceptible has received significant attention. Embedding an imperceptible, cryptographically secure signal, or watermark, is seen as a potential mechanism that may be used to prove ownership or detect tampering. While there has been a considerable amount of attention devoted to the techniques of spread-spectrum signaling for use in image and audio watermarking applications, there has only been a limited study for embedding data signals in speech. Speech is an uncharacteristically narrow band signal given the perceptual capabilities of the human hearing system. However, using speech analysis techniques, one may design an effective data signal that can be used to hide an arbitrary message in a speech signal. Also included are experiments demonstrating the subliminal channel capacity of the speech data embedding technique developed here.

1. INTRODUCTION

Watermarking is a technique for embedding a cryptographic signature into digital content for the purposes of detecting copying or alteration of the content. This is accomplished using coding techniques that hide data within the image or audio content in a manner not normally detectable. This paper focuses on an, as yet, largely unexplored aspect area of audio watermarks: speech.

For audio watermarking, Preuss, *et. al.* [8] invent a digital information hiding technique for audio using the techniques of spread spectrum modulation. Boney, *et. al.* [2] explicitly make use of MPEG-1 Psychoacoustic Model to obtain the frequency masking values to achieve good imperceptibility. Recently Riuz and Deller [9] propose a speech watermarking method for the application to the digital speech libraries. These methods have been extensively applied for music applications, but embed information over a very wide audio band based on human hearing capabilities. A potential attacker need only low-pass filter the resulting signal to remove most of the watermarking information.

Speech differs from music in their acoustic characteristics and watermarking requirements. Speech is an acoustically rich signal that it uses only a small portion of the human perceptual range. Typical speech reproduction hardware, although often the same as used with music, includes much lower bit rate channels such as telephone or compressed voice "vocoders." However, the same analysis techniques employed in such voice coding schemes can easily be adapted to create an audio watermarking signal that is robust to speech channels. Presented here is a technique for encoding an additional, arbitrary digital message into speech signals. By making use of the well understood techniques of speech analysis, significantly higher bit rates can be embedded without effecting the perceived quality of the recording.

The digital hiding technique for speech can be applied to copyright protection for digital speech libraries, audio books, as well as covert communication channel. The embedded information may be any digital message. Messages that can be used to prove authorship require the generation of an appropriate cryptographically secure digital message and are beyond the scope of this paper. However, consult [4] for information on the application of watermarks.

2. VOICEBAND SPREAD SPECTRUM SIGNAL

In contrast to previous work on audio watermarking, the speech signal is a considerably narrower bandwidth signal. The long-time-averaged power spectral density of speech indicates that the signal is confined to a range of approximately 10 Hz to 8 kHz [6]. In order that the watermark survives typical transformation of speech signals, including speech codecs, it is important that the watermark be limited to the perceptually relevant portions of the spectra. However, the watermark should remain imperceptible. Therefore, a spread-spectrum signal with an uncharacteristically narrow bandwidth will be used.

Using a direct sequence spread spectrum [3] signal, we wish to design a PN sequence with a main side lobe that fits within a typical telephone channel [5], which ranges from 250 Hz to 3800 kHz. In this work, the message sequence and the PN sequence are modulated using simple Binary Phase Shift Keying (BPSK). The center frequency of the carrier is chosen to be $f_c = 2025$Hz. The clock rate of the PN sequence, or *chip rate*, is taken to be 1775Hz, which is half of the signal bandwidth. Because the width of our watermark is very close to the modulation frequency, it is necessary to low pass filter the spread spectrum signal before modulation to prevent excessive aliasing. For this, we have chosen to use a seventh order Butterworth filter with a cutoff of 3400 Hz.

Figure 1 illustrates the power spectral density of the watermark signal, with the long-term average speech power spectrum (for both a male and female speaker) for illustration. The simplest implementation of a speech watermark system would involve adding this signal, which sounds primarily like radio static, to the speech signal at the appropriate gain. However, taking advantage of our knowledge of the speech signal itself, we are able to embed a significantly higher gain signal using techniques that are the subject of the next two sections.

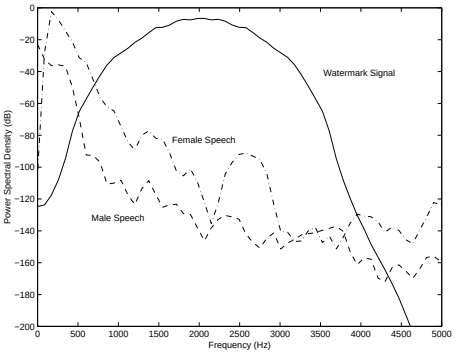

Figure 1: Power spectral densities of the watermark, male voice, and female voice.

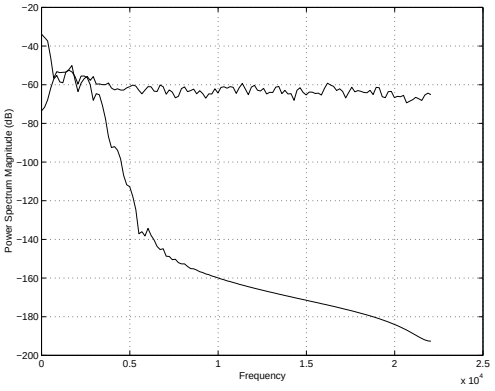

Figure 2: Power spectrum of a segment of speech and spectrum of LPC-shaped watermark signal.

3. LPC ANAYLSIS AND FILTERING

Our goal is to add as much watermark signal energy as possible to the speech signal, while still satisfying the constraint that the added signal not be perceivable when listened to. Most watermarking approaches rely on a perceptual model of human hearing. Speech is an inherently complex stimuli with rapidly changing spectral characteristics. Conventional masking effects are most often studied for spectral bands outside the range of speech, above 4 kHz. However, an effective *production model* for speech is available. The well known technique of linear prediction has proven to be highly effective in modeling speech signals. In addition, human speech perception reflects the production system characteristics. Our findings indicate that using the production model can provide excellent hiding characteristics.

In our watermark signal embedding algorithm, the watermark signal is filtered to match the overall spectral shape of the speech signal. In addition, the linear predictive analysis provides an effective dynamic measure of the degree of noise already present in the speech signal. Portions of speech that have a highly white spectrum, fricative sounds and the rapidly changing plosives sounds, are especially good candidates for embedding additional watermark energy.

Linear predicative analysis of speech involves computing the maximum likelihood coefficients of an all-pole filter of the form

$$A(z) = \frac{1}{a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}} \qquad (1)$$

There is a considerable literature on the application of linear prediction to speech signals. For our analysis, we have chosen to use the Levinson-Durbin recursive technique for evaluating LPC coefficients a_i from the short-term autocorrelation coefficients.

The short term autocorrelation can be computed from the windowed speech frame $s(t)$ as

$$r_i = \sum_{n=1}^{N-1} s(n)s(n-i)$$

which,

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix}$$

which, in vector notation can be represented by

$$\mathbf{Ra} = \mathbf{r}$$

The prediction residual energy, or the average squared-error can be computed as

$$E = \mathbf{a}'\mathbf{Ra}$$

is a measure of the "predictability" of the speech signal, and an effective measure of the noise content.

Before filtering the watermark signal using the all-pole filter, a bandwidth expansion operation is performed. This moves all of the poles closer to the center of the unit circle, increasing the bandwidth of their respective resonances. The vocal tract filter often tends to have quite narrow spectral peaks. Due to masking phenomena, sounds near these peaks are unlikely to be perceived by the listener. Therefore, by increasing the bandwidth of formant responses, larger overall watermark signal gains should be tolerable. The bandwidth parameter γ is used to adjust the LPC coefficients

$$a_i' = a_i \gamma^i$$

where γ may be chosen between 0 and 1.

Figure 2 shows the power spectrum of a segment of speech, and the spectrum of the watermark signal that results after filtering using the spectral envelope of the speech segment.

4. WATERMARK SIGNAL GAIN

The instantaneous watermark gain is dynamically determined to match the characteristics of the speech signal. In the simplest case, when little speech energy is present (i.e. during silence) the watermark is added using a fixed gain threshold. This is selected so that the watermark becomes the effective noise floor of the recording. Perceptually, a small amount of noise is always expected in a recording and the watermark signal is not atypical of such recording noise. In many applications, silence may not be transmitted or might be by coded using extreme compression. In these circumstances, designers should choose an error correcting code (such as a convolutional code) with the proper characteristics so that the message may be recovered despite these losses.

The normalized per sample speech energy E_s for one frame is $E_s = \frac{1}{N}\sum_{k=1}^{N} s^2(n) = \frac{1}{N} r_0$.

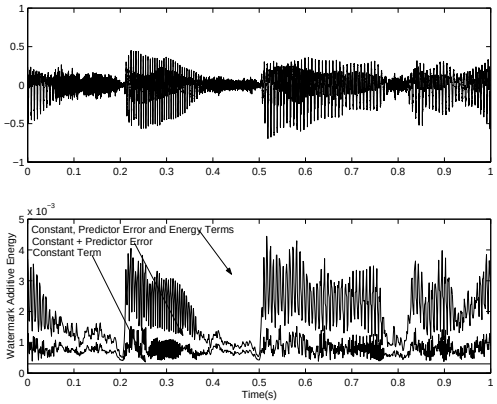

Figure 3: A segment of speech and the corresponding watermark gains.

The watermark gain in each frame can be determined by the linear combination of the gains for silence, normalized per sample residual energy E, and normalized per sample speech energy E_s,

$$g(t) = \lambda_0 + \lambda_1 E + \lambda E_s, \tag{2}$$

which is designed to maximize the strength of the watermark signals without incurring perceptual degradations. Figure 3 shows a segment of speech and the embedded watermark signal. The resulting watermarked speech is shown also in Figure 3. Listening test demonstrates that the watermarked speech is indistinguishable from the original speech with this watermark gain. If the gain is increased further, there will be "hoarseness" in the watermarked speech. Though it hardly affects the naturalness of the voice, the difference with the original speech is indeed perceptible.

5. WATERMARK DETECTION

At the receiving end, the received signal $r_0(t)$ is given by

$$r_0(t) = \sum_{t=1}^{N} w(t) + s(t) + I_0(t), \tag{3}$$

where $w(t)$ is the LPC-shaped watermark signal, $s(t)$ is the original speech signal, and $I_0(t)$ is some deliberated attacks or digital signal processing. We estimate the LPC coefficients from the received signal, and then take the inverse LPC filtering of $r_0(t)$ to get $r(t)$. After inverse LPC filtering, voiced speech becomes periodic pulses, and unvoiced speech becomes whitened noise. As is typical for speech processing, we model the inverse filtered $s(t)$ as White Gaussian Noise (WGN). Inverse LPC filtering decorrelates the speech samples $s(t)$ as well as equalizes the watermark signal $w(t)$. A correlation receiver,

$$\sum_{t=1}^{N} d(t) r(t) \overset{H_1}{\geq} 0, \tag{4}$$

gives us optimum detection performance in AWGN [7], where N is the length of a frame, in which one message bit is embedded, $d(t)$ is the despreading function, which is the synchronized, BPSK modulated spreading function for the current frame. The correlation with $d(t)$ can average out the interference, thus providing the desired robustness property. The decoding rule is a maximum likelihood decision rule, which is also a minimum probability-of-error rule since 0 and 1 in the message are sent with equal probabilities.

The problem of synchronization when the original message is not available is beyond the scope of this paper. However, the PN sequence used in the spread spectrum modulation can be used to drive a phase locked loop during decoding. The techniques presented in [3] [8] can be used in our framework for synchronization purposes.

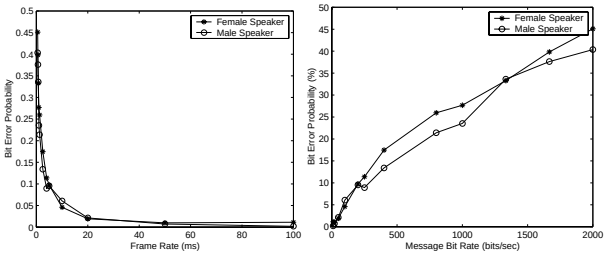

Figure 4: Bit Error Probability versus Frame Rate, and Bit Error Probability versus Message Bit Rate.

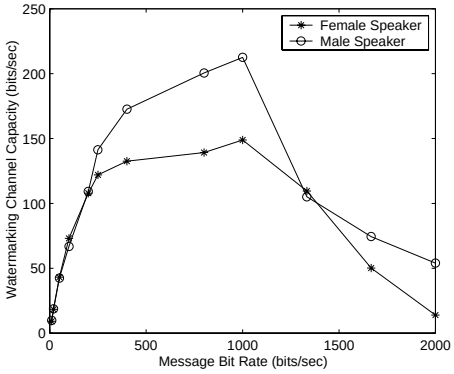

Figure 5: Watermarking Channel Capacity versus Message Bit Rate

6. EMBEDDED CHANNEL CAPACITY

A set of simulation experiments were performed to demonstrate the relationship between the frame size and message rate (1 bit per frame) and the bit error probability, as shown in Figure 4.

The spread spectrum signal, when added to the original speech, can be considered as a noisy communication channel, called the watermarking channel. The watermark is the content of the transmitted message. Without loss of generality, the message is considered to be a binary signal with equal probability for 0 and 1. The watermark channel is binary symmetric The channel capacity, which is the theoretical maximum rate for data transmission, is defined for the watermarking channel [1]:

$$C = R(1 + p log_2 p + (1-p) log_2 (1-p)), \tag{5}$$

where p is the crossover probability, R is the message bit rate. The simulation results for the watermarking channel capacity are plotted in Figure 5. For a binary symmetric channel, the chan-

Compression Scheme	Speech Bandwidth	Speech Bit Rate	Watermark Bit Reliability
16 bit linear PCM	22 kHz	706 kbps	74.05%
16 bit linear PCM	4 kHz	128 kpbs	71.58%
IMA ADPCM	4 kHz	32 kbps	68.65%
GSM 6.10	4 kHz	13 kpbs	61.23%

Table 1: Watermarking Attacks by Voice Compression

nel capacity is achievable [1]. That is, transmission codes can be designed for reliable communication under or at this rate.

The plot shows that the frame size needs to be small when high channel capacity is desired. However, the LPC prediction suffers when the frame size is too small, which makes LPC shaping less effective. And also the degradation of the watermarking channel due to attacks is more severe for smaller frame, see Section 7. Therefore, there is an intrinsic tradeoff between channel capacity and survivability of watermark. To achieve high channel capacity, good LPC predictability, and reasonable survivability simultaneously we have chosen 800 bits per second as our message embedding rate.

7. ROBUSTNESS

Watermarked media is subject to a variety of attacks. With images, images may be cropped, rotated, filtered, or otherwise changed. Audio signals are less subject to these types of manipulations, as the human perceptual system is quite sensitive to changes in audio signals. However, speech signals may be affected by transformations that include: analog to digital and digital to analog conversions, filtering, re-equalization, changes in playback rate, and compression. The algorithm presented here puts all of the watermark signal in the most perceptually important areas of the speech signal. Therefore, primitive attempts to remove the watermark by filtering are almost certain to prove ineffective.

In order to demonstrate the robustness of the data embedding scheme, we have used an analog reproduction system to simulate a crude attempt at duplication. A recording is made at 8 kHz, significantly reducing the bandwidth, and then the signal is re-sampled at the original rate. This could be considered similar to recording across a telephone channel, although no explicit telephone network equalization was applied. Finally, these 8 kHz recording were compressed and decompressed using the typical speech compression algorithms IMA ADPCM and GSM 6.10. The results are summarized in Table 7.

8. APPLICATIONS AND FUTURE WORK

This paper presents a technique for embedding an arbitrary message in a speech signal. In order to provide a complete watermarking application, one must choose a message that provides the appropriate cryptographic properties, such as proof of authenticity or ownership. In this respect, the embedding algorithm presented here can be used with nearly any comparable application. For example, it can be applied to the copyright of the language-learning CD's, audio books, recorded teleconferencing data, digital speech libraries [9] and Internet radio broadcasts, etc.

In addition, a speech data embedding algorithm suggests some new and possibly unique applications. For example, a closed captioning system can be built using the data embedding algorithm presented here, where the text transcription of the speech would be hidden in the speech itself. In addition, in-band signaling applications, typically done using dual tone "touch-tone" signals can be replaced with embedded control signals, suggesting novel simultaneous voice and data applications. For the purpose of side-information embedding, there is little threat from intentional attacks. Thus, a larger capacity of information can be communicated with less dependency on the redundancy of error correct codings.

9. REFERENCES

[1] R. E. Blahut. *Principles and Pratice of Information Theory*. Addison-Wesley Publishing Company, 1987.

[2] L. Boeny, A. H. Tewfik, and K. N. Hamdy. Digital watermarks for audio signals. In *Proc. of Multimedia 1996*, Hiroshima, 1996.

[3] G. R. Cooper and C. D. McGillem. *Modern Communications and Spred Spectrum*. McGraw-Hill Book Company, New York, 1986.

[4] F. Hartung and M. Kutter. Multimedia watermarking techniques. In *Proceedings of the IEEE*, vol. 87, July, 1999.

[5] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *ICASSP*, pages 109–112, Albuquerque, NM, 1990.

[6] N. S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1984.

[7] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, New York, 1994.

[8] R. Preuss, S. Roukos, A. Huggins, H. Gish, M. Bergamo, and P. Peterson. *Embedding Signalling*. US Patent 5319735, 1994.

[9] R. J. Ruiz and J. R. Deller. Digital watermarking of speech signals for the national gallery of the spoken word. In *ICASSP*, Turkey, 2000.