HIGH RATE DATA HIDING IN SPEECH SIGNAL

Ehsan Jahangiri and Shahrokh Ghaemmaghami
Electronics Research Center, Sharif University of Technology, Tehran, Iran

Keywords: Data Hiding, Steganography, Encryption, Multi-band Speech Coding, MELP.

Abstract: One of the main issues with data hiding algorithms is capacity of data embedding. Most of data hiding methods suffer from low capacity that could make them inappropriate in certain hiding applications. This paper presents a high capacity data hiding method that uses encryption and the multi-band speech synthesis paradigm. In this method, an encrypted covert message is embedded in the unvoiced bands of the speech signal that leads to a high data hiding capacity of tens of kbps in a typical digital voice file transmission scheme. The proposed method yields a new standpoint in design of data hiding systems in the sense of three major, basically conflicting requirements in steganography, i.e. inaudibility, robustness, and data rate. The procedures to implement the method in both basic speech synthesis systems and in the standard mixed-excitation linear prediction (MELP) vocoder are also given in detail.

1 INTRODUCTION

The modern broadband technologies have significantly improved the transmission bandwidth, which has made the multimedia signals such as video, audio and images quite popular in Internet communications. This has also increased the need for security of the media contents that has recently gained much attention. A typical approach to the issue is to provide secure channels for communicating entities through cryptographic methods. However, the use of encrypted signals over public channels could make malicious attackers aware of communications of secret messages. Such attacks may even include the attempts for disconnecting the transmission links through jamming, in the cases that the plaintext is inaccessible.

To solve the problem arising with encryption, steganography is employed that refers to the science of "invisible" communications. While cryptography conceals the secret message itself, steganography strives to hide presence of secret message from potential observers. Steganography is essentially an ancient art, first used by the Romans against the Persians, but has evolved greatly over recent years (Kharrazi et al., 2004).

A typical representation of the information hiding requirements in digital audio is the socalled magic triangle, given in Figure 1, denoting inaudibility, embedding rate, and robustness to manipulation. Basically, there is a tradeoff between these factors. For instance, by increasing the embedding rate, inaudibility may be violated. This is particularly more critical in audio, as compared to image or video, because Human Auditory System (HAS) is more sensitive to deterioration or manipulation than the human visual system (Agaian et al., 2005). This means that embedding rate in secure transmission of audio signals is more challenging than that of digital images.

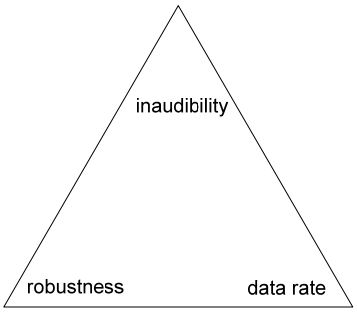

Figure 1: Magic triangle for data hiding.

Most steganography techniques proposed in the literature use either psychoacoustic properties of HAS, such as temporal and spectral masking properties (Gopalan and Wenndt, 2006), or spread spectrum concepts (Matsuoka, 2006). Gopalan and Wenndt (Gopalan and Wenndt, 2006) employed spectral masking property of HAS in audio

Jahangiri E. and Ghaemmaghami S. (2007).
HIGH RATE DATA HIDING IN SPEECH SIGNAL.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 287-292
DOI: 10.5220/0002137102870292
Copyright © SciTePress

steganography. They used four tones masked in frames of the cover signal and, based on relative power of this tones, achieved an embedding capacity of two bits per frame that led to a maximum embedding capacity of 250 bps. In (Gopalan, 2005) Gopalan used the same strategy as that in (Gopalan and Wenndt, 2006) in cepstrum domain.

Chang and Yu (Chang and Yu, 2002) embedded covert message in the final stage of multistage vector quantization (MVQ) of the cover signal. Because most of signal's data is extracted in the primary stages, embedding data in the last stage makes no substantial perceptual difference. They could embed 266.67 bps in the four-stage VQ of Mixed Excitation Linear Prediction (MELP) speech coding system, introduced by (McCree et al., 1997), and 500 bps in the two-stage VQ of G.729 standard coder (ITU 1996). The phase coding technique proposed in (Bender et al., 1996) could embed only 16-32 bps. The echo-based coding algorithm (Mansour, 2001) achieved an embedding rate of about 40-50 bps. Ansari et al. (Ansari et al., 2004) claimed reaching a capacity of 1000 bits of data in a one-second segment of audio, using a frequency-selective phase alteration technique.

In this paper, we propose a different approach to high capacity data hiding that can embed a large amount of encrypted message in unvoiced parts of speech signals conveyed by a typical voice file, e.g. a wav file. The proposed method exploits the noise-like signal, resulting from a data encryption process, to construct unvoiced parts of speech signal in either a binary or a multi-band speech synthesizer.

The rest of paper is organized as follows. The main concept of the proposed method is described in section 2 and basic implementation of the method is given in section 3. Section 4 addresses the multi-band based implementation and section 5 is allocated to implementation of the method in the MELP coding model. The paper is concluded in section 6.

2 THE PROPOSED METHOD

The key idea in the proposed method for increasing the hidden data embedding capacity is to exploit a voicing-discriminative speech synthesizer, within a high-capacity voice filing framework, to generate cover signal. This releases a large data space in the voice file that is used to accommodate the encrypted covert message. The simplest structure for such a speech synthesis system uses a binary excitation model, in which each frame of the signal is reconstructed by applying either a periodic pulse train (for voiced speech) or a random sequence (for unvoiced speech) to the synthesis filter (Chu, 2003). In this basic coding scheme, the covert message is converted into a noise-like sequence through encryption, which is employed instead of a random generator to excite the synthesis filter to produce unvoiced frames.

The encrypting process attempts to remove correlation between samples and makes ciphertext a noise-like sequence. This can be achieved by using a stream cipher, for instance, in which the ciphertext is obtained from a simple function of exclusive-or between plaintext stream and key stream ($C=P \oplus K$; Figure 2). In the simplest form, Linear Feedback Shift Registers (LFSRs), which satisfy Golomb's criteria in one period, can be used to generate the key stream (Beker and Piper, 1982). It can be shown that if any bit of key stream occurs independently with occurrence probability of 0.5 for bits 0 and 1, the ciphertext is also an independent identically distributed (i.i.d) stream with occurrence probability of 0.5 for bits 0 and 1.

There are some tests proposed by National Institute of Standards and Technology (NIST) in order to determine randomness degree of a stream cipher. Any stream cipher of a higher degree of randomness can be more secure. Stream ciphers like SNOW.2 and SOSEMANUK, both with key sizes of 128 or 256 bits, can provide adequate degree of randomness.

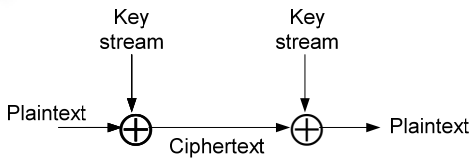

Figure 2: Stream cipher scheme.

Figure 3 shows block diagram of the basic embedding method using a simple binary excitation speech synthesizer. The hiding process is reversible, such that the ciphertext can be extracted from the cover signal at the decoder and is deciphered to attain the covert message. In order to exactly recover original plaintext, it is required to employ an error-free encryption method associated with a reliable extraction process. Using stream cipher in encryption comes with the advantage of avoiding error propagation that is of great concern here. This is because encryption in stream cipher is a bit-wise process with no feedback loops. Conversely, in block ciphers, like AES, a flipped bit could affect

the whole ciphertext or the reproduced plaintext depending on the mode of use (see Heys, 2001).

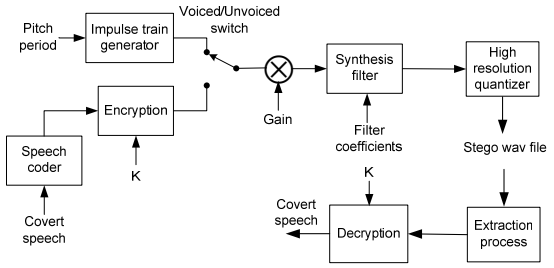

Figure 3: Block diagram of the basic embedding method.

To achieve a higher performance in both cover signal quality and information hiding capacity, we use a multi-band excitation (MBE) speech coding system (Griffin and Lim, 1988), rather than the binary excitation model mentioned earlier. The MBE based vocoders resolve the complexity associated with "mixed" voiced/unvoiced characteristics of speech. Some speech coding algorithms based on MBE, such as INMARSAT-M (Kondoz, 1994) and MELP (McCree et al., 1997), substantially improve quality of the synthetic speech, as compared to non-MBE vocoders in low bit rates.

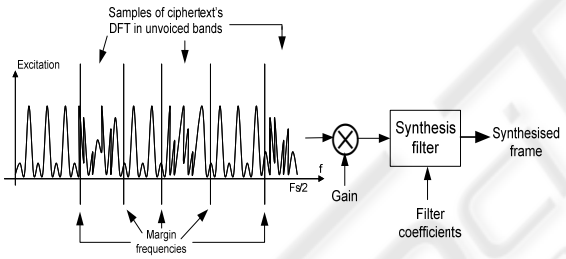

Figure 4: Data hiding based on multi-band speech synthesis.

In an MBE coder, the excitation spectrum is taken as a series of voiced/unvoiced (v/uv) bands that are computed and arranged based on the original signal spectrum for each frame of the signal (Chiu and Ching, 1994). This allows each speech segment to be partially voiced and partially unvoiced in the frequency domain. Although there is basically no limits to the number and patterns of v/uv bands, it has been shown in (Chiu and Ching, 1994) that a small number of v/uv bands can adequately reconstruct a near natural and intelligible speech signal. Many other findings in low-rate speech coding confirmed this assertion (see e.g. McCree et al., 1997).

Figure 4 illustrates an MBE based speech synthesis system. We replace the excitation signals, in unvoiced bands, with the ciphertext that conveys the covert message. The embedding procedure is reversible, such that the message can be recovered from the synthesized speech by an authorized receiver. More details are given in the next sections.

3 IMPLEMENTATION

The procedure described here uses a binary excitation model in an LPC (Linear Prediction Coding) system to generate the cover speech, as shown in figure 3. This is a simple, basic structure for implementing the method, in which a frame of speech is assumed to be fully periodic (voiced) or entirely noise-like (unvoiced). In this basic experimental model, the signal is sampled at 8 kHz and is decomposed into 40ms frames (320 samples), with 50% overlap, using a rectangular window. The whole excitation sequence is then reconstructed through an overlap-add procedure applied to all voiced and unvoiced frames. This long excitation sequence, of the same length of cover signal, contains noise-like parts that are replaced by the ciphertext of covert message.

The resulting excitation sequence in then segmented into the same frame lengths, with 50% overlap, which excite the synthesis filter constructed using the coefficients calculated in the LPC analysis. The cover signal, now containing the ciphertext, is generated by an overlap-add procedure applied to the synthesized speech at the output of the synthesis filter. The resulting speech, called the stego signal (sounds like cover), can be located in a typical voice file, e.g. in wav format. It is to be noted that the ciphertext remains detectable if the excitation signal is constructed with an error less than one-half of the quantization step in unvoiced frames.

The ciphertext detection process uses inverse filtering in the LPC model to retrieve the excitation signal. However, because one-half of successive excitation frames are identical in our basic 50% overlapped framing procedure, same parts are used to excite j^{th} and $(j-1)^{th}$ synthesis filters in the first half of the j^{th} frame of synthesized stego speech, over the range of $(j-1)\times 160 \le n < j\times 160$ (n is the sample index). Hence, we can extract the first half of the j^{th} frame of excitation by inverse filtering of the stego signal in this interval, using $(j-1)^{th}$ and j^{th} synthesis filters, as:

$$H_{j^{th}}(z) = \left[\frac{g(j-1)}{1-\sum_{i=1}^{10} a_{(j-1),i} z^{-i}} + \frac{g(j)}{1-\sum_{i=1}^{10} a_{j,i} z^{-i}}\right]^{-1} \quad (1)$$

where $g(0)=0$, $a_{0,i}=0$, and $a_{(j-1),i}$, $a_{j,i}$, $g(j-1)$, and $g(j)$ are coefficients and gains of $(j-1)^{th}$ and j^{th} LPC synthesis filters, respectively.

The above-mentioned procedure can precisely recover the excitation sequence. However, due to finite register length of calculations in the employed implementation platform, some errors may be encountered. For instance, in MATLAB (64-bit floating-point), the error between original and the extracted excitation sequence is bounded by something less than 0.5×10^{-12}. This error determines the number of bits that we can allocate to each sample of excitation signal in unvoiced frames, which is calculated as:

$$0.5\times10^{-12} < \Delta / 2 = X_m / ((2^n-1)\times2) \qquad (2)$$

where Δ is the quantization step for unvoiced excitation samples, X_m is the quantization range that is 1 here, and n is the number of bits per sample.

In no-quantization case of stego speech, we can allocate at most 39 bits to each sample of excitation in unvoiced parts. In a practical system, however, we need to quantize the synthesized stego speech that restricts us to lower number of bits allocated to each unvoiced sample. In this basic experiment, we can use a 16 or 32 bits per sample PCM signal in 'data' chunk of a wav format file. As an actual example, assuming that 25% of speech frames are unvoiced, and allocating 8 bits per sample to unvoiced excitation, we reach an embedding rate of 0.25×8×8kHz=16 kbps, in 8kHz sampling rate.

4 MBE BASED HIDING

The basic embedding procedure, described earlier, can be applied to an MBE based speech coding system that discriminates periodic and noise-like components of the signal in individual bands in frequency domain. To demonstrate the method in such a paradigm, we use a simple dual-band speech synthesizer as the simplest MBE structure. It has been shown that for the case of two excitation bands, the lower frequency one is usually voiced while the other is unvoiced. Thus in our MBE implementation we embed covert message in upper frequency band. An example of such a dual-band speech synthesis system can be found in (Chiu and Ching, 1994). Cover signal, sampling rate, overlap percentage and frame length are all the same as those used in the binary excitation experiment. However, unlike the binary excitation system, in which we embedded ciphertext in time domain, we embed DFT (Discrete Fourier Transform) of ciphertext in unvoiced bands in frequency domain. This is due to the MBE model that is typically implemented in frequency domain.

The ciphertext is embedded in every other frames, e.g. odd frames, to avoid ciphertext muddle due to overlapping structure (50% overlap in this case), where deterministic random sequences are used to form unvoiced bands of even frames. Hence, an authorized receiver can generate even frames, reconstruct odd frames, and then extract corresponding ciphertext from unvoiced bands, by removing the overlapping effect. Average embedding capacity in this system depends on the mean of voiced/unvoiced transition frequency, which we found to be about 2.2 kHz in a typical 4 kHz 2-band excitation system (Figure 5). This leads to an embedding rate of 28.8 kbps for 8 bits per sample encoding schemes.

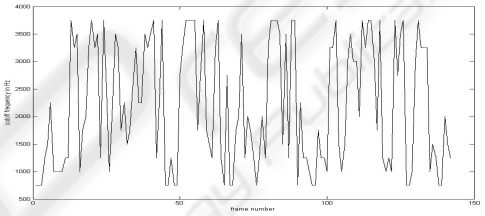

Figure 5: Demonstration of margin frequencies between voiced and unvoiced bands for frames of cover speech.

In general, the embedding capacity is given as:

$$C = (f_s - 2\times \overline{f_m})\times BPS \qquad (3)$$

where f_s, $\overline{f_m}$, and BPS are sampling frequency, average transition frequency in cover speech, and the number of bits per sample, respectively.

This MBE based scheme can be generalized for most MBE based speech synthesis systems with any overlapping structure. The use of the proposed method in a standard MBE based coding system is described in the next section.

5 DATA HIDING IN MELP

A block diagram of the MELP model of speech production is shown in Figure 6. Periodic excitation and noisy excitation are first filtered using the pulse shaping filter and noise shaping filter, respectively. Signals at the filters' outputs are added together to form the "mixed" excitation. In FS MELP (McCree et al., 1997), each shaping filter is composed of five 31-tap FIR filters, called the synthesis filters, which are employed to synthesize the mixed excitation signal in the decoding process. Each synthesis filter

controls one particular frequency band, with pass-bands assigned as 0–500, 500–1000, 1000–2000, 2000–3000, and 3000–4000 Hz. The synthesis filters, connected in parallel, define the frequency responses of the shaping filters. Responses of these filters are controlled by a set of parameters called voicing strengths; these parameters are estimated from the input signal.

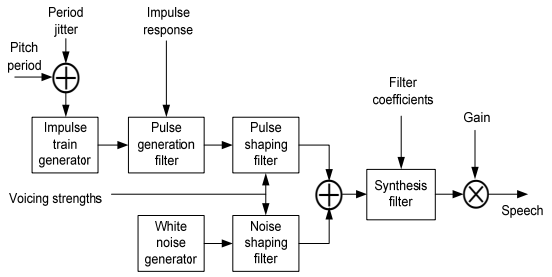

Figure 6: The MELP model of speech production (reproduced from (Chu, 2003)).

By varying the voicing strengths with time, a pair of time-varying filters results. These filters decide the amount of pulse and the amount of noise in the excitation at various frequency bands (Chu, 2003). Denoting the impulse responses of the synthesis filters by $h_i[n],\ i = 1\ to\ 5$, the total response of the pulse shaping filter is:

$$h_p[n] = \sum_{i=1}^{5} vs_i h_i[n] \tag{4}$$

with $0 \le vs_i \le 1$ being the voicing strengths. The noise shaping filter, on the other hand, has the response:

$$h_n[n] = \sum_{i=1}^{5} (1 - vs_i) h_i[n] \tag{5}$$

Thus, the two filters complement each other in the sense of the gain in frequency domain. Normalized autocorrelation and aperiodic flag determines voicing strength of each band, which is quantized with one bit per frame. During decoding for speech synthesis, the excitation signal is generated on a pitch-period basis, where voicing strengths are linearly interpolated between two successive frames. Thus, even though transmitted voicing strengths in coded bit stream is 0 or 1 for each frame, they have values in the interval [0,1] between two frames during interpolation at the decoder side.

In order to achieve a reversible embedding process, generation of the pulse excitation (shown in the upper branch of mixed excitation in Figure 6) should be repeatable at the decoder by an authorized receiver. Following encryption of covert message, the DFT of ciphertext is embedded in unvoiced excitation bands ($vs \le 0.6$), where each unvoiced band is multiplied by complement voicing strength ($1 - vs$) of that band. By adding noise excitation, that includes the DFT of ciphertext, to the pulse excitation, we construct a mixed excitation signal to excite the synthesis filter.

The only random variable in the pulse excitation is the period jitter (see Figure 6) that is usually distributed uniformly over the range of ±25% of the pitch period to generate erratic periods, simulating the conditions encountered in transition frames. The actual pitch period to use is given as:

$$T = T_0 (1 + jitter.x) \tag{6}$$

where T_0 denotes the decoded and interpolated pitch period, and x represents a uniformly distributed random number in the interval [-1, 1]. For voiced frames, the value of jitter is assigned according to jitter←0.25, if aperiodic flag is equal to one; otherwise, jitter←0 (Chu, 2003). Thus, in order to build a pulse excitation to be reproducible at the authorized decoder, we generate a random but deterministic x uniformly distributed over the interval [-1,1]. This deterministic random sequence can be the key stream of a stream cipher that the authorized decoder has its initial key.

To attain ciphertext, we produce mixed excitation by filtering the synthesized stego speech by inverse filter of spectral enhancement filter, pulse dispersion filter, and synthesis filter in cascade. Subsequently, the mixed excitation is subtracted from the pulse excitation signal, generated at the authorized decoder side, to get noise excitation signal that includes DFT of the ciphertext. Then, we multiply unvoiced bands of the noise excitation signal by inverse of related complement voicing strengths to extract the DFT of ciphertext, which is then computed using inverse DFT.

In generation of pulse excitation, we use 31-tap FIR filters but, for generating the noisy excitation signal, we embed the DFT of ciphertext in determined frequency intervals, using a flat frequency response filter, to make the ciphertext detectable at the authorized receiver. By using this embedding method and allocating 8 bits to each sample of noisy excitation, it is possible to embed approximately 20 kbps in a phonetically-balanced TIMIT phrase as cover speech.

It is to be noted that, unlike most typical stenography methods, there is no simple tradeoff between embedding capacity, inaudibility, and the quality of reconstructed speech in the proposed method. Rather, structure of the coding system and the multi-band excitation scheme designate

interrelation between these attributes. This is while inaudibility is always guaranteed, if no statistical restrictions are imposed on the pseudo-random sequences employed to generate unvoiced bands.

6 CONCLUSIONS

In this paper, we have introduced a novel method for hiding data in a cover voice file that can yield a high data embedding rate. In this method, an encrypted covert message is embedded in the unvoiced bands of speech signal, encoded by an MBE-based coding system, which leads to a high data hiding capacity of tens of kbps in a typical digital voice file transmission scheme. By using this method, it is possible to embed even a larger than the host covert message within the cover signal. The method also provides an unsuspicious environment for data hiding strategies, e.g. steganography, due to keeping the statistical properties of the cover speech almost unchanged. However, the ultimate chance for an attack to the system to detect the message will remain the same as that in a cipher system used to encrypt a secret message.

REFERENCES

Agaian, S.S., Akopian, D., Caglayan, O., D'Souza, S.A., 2005. Lossless Adaptive Digital Audio Steganography. *Thirty-Ninth Asilomar Conference on Signals, Systems and Computers, October 28 - November 1, On page(s): 903- 906.*

Ansari, R., Malik, H., Khokhar, A., 2004. Data-hiding in audio using frequency-selective phase alteration. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), 17-21 May, vol.5 on page(s): V- 389-92.*

Beker, H., and Piper, F., 1982. *Cipher Systems: The Protection of Communications,* John Wiley & Sons.

Bender, W., Gruhl, D., Morimoto, N., Lu, A., 1996. Techniques for data hiding. *IBM system Journal, vol.35,nr. 3/4.*

Chang, P.C., Yu, H.M., 2002. Dither-like data hiding in multistage vector quantization of MELP and G.729 speech coding. *Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 3-6 November, Volume 2, Page(s):1199 – 1203.*

Chiu, K.M., Ching, P.C., 1994. A dual-band excitation LSP codec for very low bit rate transmission. *International Symposium on Speech, Image Processing and Neural Networks, 13-16 April, vol.2 on page(s): 479-482.*

Chu, W. C., 2003. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders,* John Wiley & Sons.

Gopalan, K., 2005. Audio steganography by cepstrum modification. *(ICASSP '05) Volume 5, 18-23 March, v/484 Vol. Page(s):v/481.*

Gopalan, K., Wenndt, S., 2006. Audio Steganography for Covert Data Transmission by Imperceptible Tone Insertion. *IASTED Con.f. Comm. Systems and Applications Banff, Alberta, Canada July 3-5.*

Griffin, D.W., Lim, J.S., 1988. Multi-band excitation vocoder. *IEEE Trans. ASSP, 36(8); August, 664-678.*

Heys, H.M., 2001. An Analysis of the Statistical Self-Synchronization of Stream Ciphers. *Proceedings of INFOCOM, Anchorage, Alaska, Apri,, pp. 897-904.*

ITU 1996. *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)—ITU-T Recommendation G.729.*

Kharrazi, M., Sencar, H.T., Memon, N., 2004. Image Steganography: Concepts and Practice. *April 22, WSPC/Lecture Note Series. http://www.ims.nus.edu.sg/preprints/2004-25.pdf.*

Kondoz, A.M., 1994. *Digital Speech: coding for Low Bit Rate Communications Systems*, John Wiley & Sons.

Mansour, M.F., Tewfik, A.H., 2001. time-scale invariant audio data embedding. *IEEE International conference on Multimedia and Expo, ICME, Japan, August.*

Matsuoka H., 2006. Spread Spectrum Audio Steganography Using Sub-band Phase Shifting. *Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP '06. on Dec, Page(s):3 – 6.*

McCree, A.V., Supplee, L.M., Cohn, R.P., Collura, J.S., 1997. MELP: The New Federal Standard at 2400 bps. *IEEE ICASSP, pp. 1591–1594.*

Sencar, H., Ramkumar, M., Akansu, A., 2004. *Data Hiding Fundamentals and Applications: Content Security in Digital Multimedia,* ELSIVIER ACADEMIC PRESS.