

Covert Speech Communication Via Cover Speech By Tone Insertion¹

Kaliappan Gopalan¹, Stanley Wenndt², Andrew Noga², Darren Haddad², and Scott Adams²

¹ Department of Engineering, Purdue University Calumet, Hammond, IN 46323
219-989-2685

gopalan@calumet.purdue.edu

² Multi-Sensor Exploitation Branch, AFRL/IFEC, Rome, NY 13441
315-330-7244

{Stanley.Wenndt, Andrew.Noga, haddadd, Scott.Adams}@rl.af.mil

Abstract --This paper presents the results of embedding a covert audio message in a cover audio signal for battlefield communication using steganography. Representing the covert message in a compressed or coded form, typically in the standard Global System for Mobile communication half-rate code (GSM 06.20), tones are added to the cover utterance in accordance with the coded bits. Based on the psychoacoustical masking property of the human auditory system, the added tones are set at very low power levels relative to the cover audio frames. The low power levels ensure that the tones are inaudible in the message-embedded stego signal. Besides imperceptibility in hearing, the spectrogram of the stego signal also conceals the existence of embedded information. Both of these features render the detection of embedding in the stego signal difficult to accomplish. Oblivious detection of the stego signal, instead of escrow detection, yields the embedded information accurately. By incorporating spread-spectrum techniques, the hidden message can be made further secure from unauthorized detection and modification. While the payload capacity of the proposed technique is low, the embedded information may be robust for retaining data under attack.

TABLE OF CONTENTS

.....	
1. INTRODUCTION.....	1
2. DATA EMBEDDING BY TONE INSERTION...2	2
3. EXPERIMENTAL RESULTS.....2	2
4. DISCUSSION.....4	4
5. CONCLUSION..... 5	5
REFERENCES	5

1. INTRODUCTION

Communication security is essential for transmitting vital information over publicly available media. Measures and controls are taken to deny unauthorized persons from intercepting information transmitted in any form – text, data, voice, image or video – and to ensure the authenticity of such communications. These measures, known under the category of COMSEC or communications security, include cryptosecurity, transmission security, emission security, and physical security of COMSEC material. COMSEC and, in general, TRANSEC (transmission security) address the problem of wireless network security, typically using a data spreading technique that continually changes the pseudo-random spread sequence to help prevent eavesdropping.

Covert speech communication is concerned with transmitting vital information by speech via an innocuous cover speech in a secure and robust manner. While a text message, or the text equivalent of a covert message, can be embedded in a cover speech, important non-speech information such as the emotional state, accent, etc. cannot be conveyed adequately by the covert text message.

Covert speech communication by embedding the message in a cover medium is an application of the art and science of steganography, or data embedding, that has been increasingly gaining importance in the all-encompassing field of information technology. While cryptography conceals the information contents being transmitted, steganography conceals the existence of covert information in the cover medium, be it audio, image, or video. In encryption, the message audio signal, for instance, is itself altered in such a way that it renders the resulting data

¹0-7803-7651-X/03/\$17.00 © 2003 IEEE
IEEEAC paper #1027, Updated November, 2002

unintelligible. Although persons without the encryption key cannot decipher the signal, transmitting encrypted information, in general, arouses suspicion about the presence of hidden information. For battlefield communication, in particular, hiding the existence of information is, therefore, crucial. Using a host medium as a wrapper or carrier in steganography, the covert information is kept intact as opposed to modifying it in cryptography.

Steganography, in general, relies on the imperfection of the human auditory and visual systems. Image and video steganography exploit the low visual sensitivity in perceiving changes in luminance of greater than one in 30 of random patterns, or one in 240 in uniform levels of gray, for example [1]. Audio steganography takes advantage of the psychoacoustical masking phenomenon of the human auditory system [HAS]. Psychoacoustical, or auditory, masking is a perceptual property of the HAS in which the presence of a strong tone renders a weaker tone in its temporal or spectral neighborhood imperceptible [2]. This property arises because of the low differential range of the HAS even though the dynamic range covers 80 dB below ambient level [2]. In temporal masking, a faint tone becomes undetected when it appears immediately before or after a strong tone. Frequency masking occurs when human ear cannot perceive frequencies at lower power level if these frequencies are present in the vicinity of tone- or noise-like frequencies at higher level. Additionally, a weak pure tone is masked by wide-band noise if the tone occurs within a critical band. We must note that the masked sound becomes inaudible in the presence of another louder sound; the masked sound, faint it may be, is still present, however. This property of inaudibility of weaker sounds is used in different ways for embedding information. In the case of embedding in phase or amplitude, for example, the phase or amplitude of a frequency-masked sample in the spectral domain is altered in accordance with the information bit to be embedded [3-5]. Instead of modifying the host sample, the present work inserts tones at low power to conceal information. The technique and the experimental results are described in the following sections.

2. DATA EMBEDDING BY TONE INSERTION

Tone insertion experiments were conducted using utterances from the TIMIT (Texas Instruments Massachusetts Institute of Technology) database as host samples. For the covert messages, (a) a random set of approximately 600 to 4000 bits, and (b) a short utterance, "seven one" spoken by a male speaker and represented compactly using the Global System for Mobile communication half-rate code (GSM 06.20) coding scheme were used.

In the first experiment, two tones at frequencies of 1875 Hz (f_0) and 2625 Hz (f_1) were generated for embedding bit 0 and bit 1 respectively. The host utterance was obtained by concatenating two utterances from the TIMIT database to get a total length of 172057 samples. Using non-overlapping segments of 256 samples (16 ms) each, the host

was divided into 672 frames. (We note that at 256-point DFT, the tone frequencies correspond to frequency indices of 31 and 43, respectively.) A random set of data bits, one for each frame, was generated. The amplitudes of the spectral components in each frame at the two tone frequencies were set in accordance with the bit to be embedded in a given frame. To embed bit 0, the amplitude of the tone at the lower frequency of $f_0 = 1875$ Hz was set to a value such that the power of the tone corresponded to 0.25 percent of the average power of the frame under consideration. In addition, the tone at the other frequency was set to a negligible value – typically to one-thousandth to one-hundredth the amplitude of the 'dominant' tone. If the bit was 1, the amplitudes of the tones were reversed; that is, the higher frequency tone ($f_1 = 2625$ Hz) was set to 0.25 percent power of the frame while the lower frequency tone was put at a negligible power. The simultaneous adjustment of significant (0.25 percent) and extremely low powers to the tones helps in two ways. First, it avoids one or both of the tones being detected in hearing – if only one of the tones is set to a fixed power ratio relative to the frame power, the other tone may be heard in some cases where the host frame inherently has a substantial component at the tone frequency. The second advantage is that a known high/low ratio of power between the tones facilitates the detection of the embedded bit even when the embedded amplitudes are scaled or quantized. The frames with their spectral components at the tone frequencies set in accordance with the data bits constituted the stego signal. For transmission, the frame-embedded signal was quantized to 16 bits, the same as the original pulse-code modulated (PCM) host audio signal. Embedded data were retrieved at the receiver from the power ratio of the spectral components at the tone frequencies f_0 and f_1 .

The second experiment extended the technique to double the payload capacity by using four tones. For this experiment, two longer TIMIT utterances were concatenated to obtain a long host for embedding a large size of data. Only one of the four tones was set to 0.25 percent of the average power of each frame while the others were set to negligible values. For detection, the ratio of frame power to power at each tone was used. This ratio, clearly, is a minimum for the tone that was set at 0.25 percent of the frame power. The results of these experiments are discussed in the next section.

3. EXPERIMENTAL RESULTS

In the first experiment with random data, a total of 672 bits were embedded in each of the 672 non-overlapping frames of two concatenated TIMIT utterances. Although each of the host utterances came from a different female speaker, there was no loss of perceptual quality in the stego signal. The two spectral components at 1875 Hz (f_0) and 2625 Hz (f_1) were either absent or weak in the unembedded host; hence, modification of the two components in all the host frames did not affect the spectrogram or hearing of the embedded audio. At the receiver, all the 672 embedded data

bits were recovered correctly from the ratios of frame power to power at the tone frequencies.

Fig. 1 shows the spectrograms of the host and the stego signals. Although every frame has the two spectral components at f_0 and f_1 – one at 0.25 percent and the other at one thousandth of frame power – regardless of the other components, they are not detectable in the spectrograms.

Thus the stego signal is not only indistinguishable in perception, it also conceals the existence of the embedded information. Consequently, existing techniques of steganalysis can not detect the presence of, much less extract, the embedded data even with the availability of the original host signal. A recently patented technique that can display spectrograms at different user-controlled

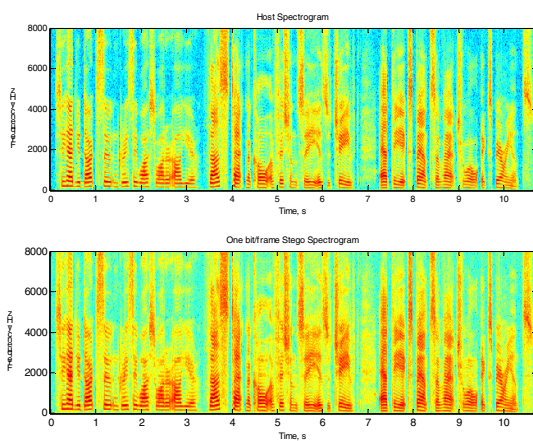


Fig. 1 Spectrograms of host (top) and stego with 1 bit/frame of random data (bottom)

frequency- and time-resolutions, for example, was unable to differentiate between the host and stego [6].

In the second experiment, four tones were used to embed two bits in each frame. In addition, successive frames for embedding were overlapped with 50 percent to further increase the payload capacity. After verifying the imperceptibility of and the data recovery from the stego signal, the technique was extended for use in covert battlefield communication in which the hidden information can be another utterance. For initial studies, the utterance, “seven one” spoken by a male speaker, was used as the covert message. This utterance, in its PCM form, has 4090 samples of 16 bits each, giving a bit capacity of 65440 bits. Even with two TIMIT utterances concatenated (resulting in a total host length of 197578 samples), the host was not long enough to accommodate the entire covert utterance with 256 samples per frame. Instead of choosing a much longer host, the covert message was represented in the GSM format resulting in a compact form of 2800 bits. Although the receiver, after extracting the bits, needs to reconstruct the utterance back, the extra effort contributes to added security in transmission while reducing the size of hidden data. With two bits inserted in each host frame of 256 samples, only

1400 overlapped frames, or 17938 samples were needed for embedding all the covert message bits.

Tones were selected at DFT indices of 11, 19, 29, and 41 in a 256-point DFT, corresponding to frequencies of 687.5 Hz, 1187.5 Hz, 1812.5 Hz, and 2562.5 Hz. Here again, the frequencies selected for insertion were either absent or weak in the host frames. With four tones, however, an additional step was necessitated to prevent the detection of embedding. Presence of a continuous stream of 0’s or 1’s in the covert data, for instance, results in the same tone being set at 0.25 percent of the corresponding frame power. Although a listener may not be able to perceive the tone because of its low power, the spectrogram is likely to show ‘holes’ at the remaining three tone frequencies. As an example, Fig. 2 shows the spectrograms of a host and the stego signal embedded with a string of 0’s in the middle half of the host while the other parts have random binary data. The frequency voids or holes can be seen at 1250 Hz, 1875 Hz, and 2500 Hz which correspond to the tones set at negligible power while the tone at 625 Hz, corresponding to the data pair [0 0], can be barely noticed. To a malicious attacker, these artifacts of frequencies are indicative of host manipulation even without the knowledge of host spectrogram. To avoid such an obvious detection of embedding, a binary key of the same size as the size of data to embed was used for each successive pair of data bits. A pair of bits from the key determined which of the four tones

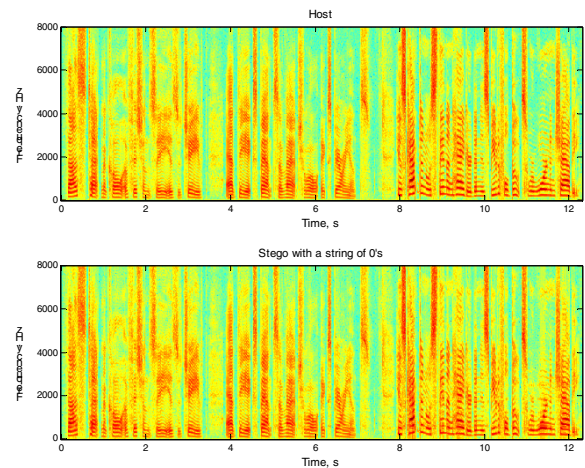


Fig. 2 Detectability of embedding - Spectrograms of host (top) and stego with a string of 0’s embedded in the middle half (bottom)

was to be set at 0.25 percent of current frame power while the others were set at negligible power. We note that each successive pair of key bits sets the order of the four tones with the one for the 0.25 percent power as the first. (To reduce the size of the key, we may use a smaller key and repeat the tone order. Depending on the data, this may still result in detectable frequency holes in the spectrogram of the stego signal.) Using the same key at the receiver, the dominant tone frequency and the order of the other three

tones were first established. The minimum of the ratio of the frame power to tone powers, along with this order, was used to determine the embedded bit pair.

The frequency-hopping scheme for the four tones was applied for embedding the GSM-encoded covert speech. The resulting spectrograms of the host and stego are shown in Fig. 3. Neither the spectrogram nor the perceptibility of the stego appeared to contain any difference from that of the unembedded host.

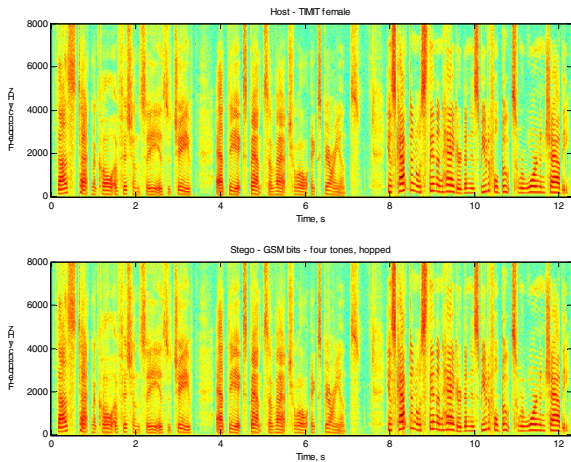


Fig. 3 Spectrograms of host (top) and stego with GSM-encoded covert utterance

In addition to the clean host from the TIMIT database, the frequency-hopped tone insertion was also used in an experiment with a noisy host from the Greenflag database. Obtained as 16-bit PCM data at a rate of 8000 samples per second, the Greenflag database consists of utterances from the cockpit of fighter aircraft. Because of the high level of noise inherent in the host, externally introduced tone or noise arising from embedding is generally not noticeable. Figs. 4 and 5 show the result of embedding the GSM-coded covert speech, ‘seven one’ on a host consisting of two utterances from the Greenflag database. Using 128 samples per frame – because of the lower sampling rate – the host of 80128 samples has 1251 frames which can embed only 2501 bits out of the 2800 bits of the coded covert speech.

Because of the high level of intrinsic noise in the host, the dominant tone power was raised to more than 10 percent. Although the stego signal did not show any perceptual difference from the host, the higher tone power started showing up in the spectrogram. To mask the dominant tone in the spectrogram, the tones were set to frequencies in the range where the host has significant energy. In the 400 Hz to 1000 Hz range, for example, the host has relatively high spectral energy over almost the entire duration. Hence, inserting tones at frequencies of 562.5 Hz, 687.5 Hz, 812.5 Hz and 1000 Hz, with as much as 25 percent of frame

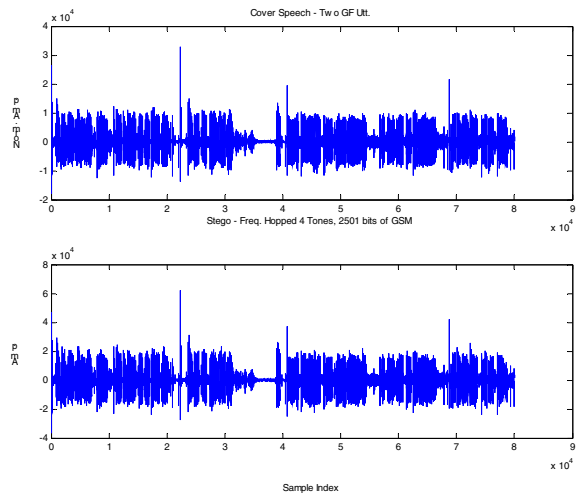


Fig. 4 Greenflag utterance host (top) and GSM-coded bits embedded stego

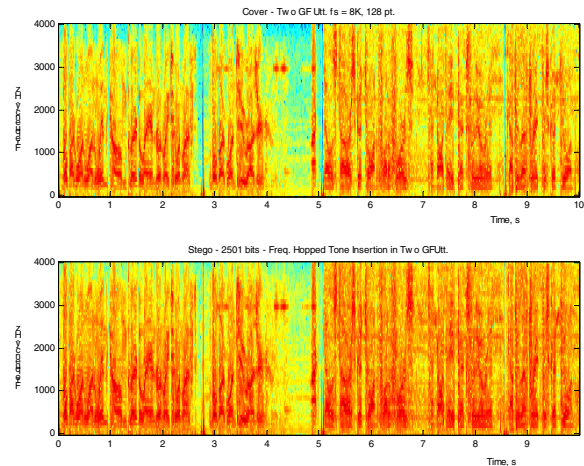


Fig. 5 Spectrograms of Greenflag host (top) and stego with 2501 bits (bottom)

power in the dominant tone did not result in any noticeable difference in speech quality or spectrogram; the inserted tones, randomized because of the hopping key, were clearly masked by the already significant spectral components in the host. Fig. 6 shows the spectrograms of the host and stego for comparison.

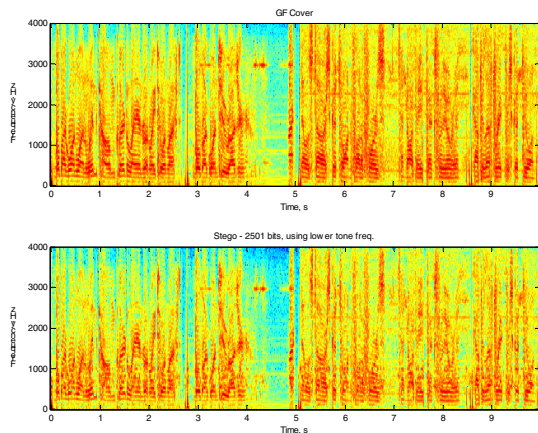


Fig. 6 Spectrograms of Greenflag host (top) and stego with 2501 bits with stronger tones at lower frequencies (bottom)

4. DISCUSSION

The results of the present experiments clearly demonstrate the feasibility of the proposed technique. For general covert speech communication using any given host speech, however, several points must be considered. These are discussed briefly next.

Choice of Tone Frequencies

The results of the experiments indicate that at low power the tones may be set to any two (for one-bit embedding) or four (for two-bit embedding) frequencies. To avoid loss of data due to filtering of high frequency noise in the received stego, for example, it is preferable to select the tones in a midband. Any attack on the stego in an attempt to destroy the embedded information cannot succeed without destroying the ‘cover’ audio as well. Tones inserted at lower frequencies, particularly in the vicinity of fundamental frequency F_0 , may affect the speech quality. If the tones at lower energies are in close proximity to F_0 , the stego may be perceptually different from the host, especially if the host is clean. Low power tones in the neighborhood of formants are not likely to affect the speech quality.

For noisy host utterances, tones can be set at higher power levels and yet be indistinguishable in hearing. However, as pointed out in the previous section, they may be noticeable in the stego spectrogram. By selecting tones from high energy regions of the host, they can be masked in the hearing and the spectrogram. Clearly, the choice of tones needs to be made carefully for imperceptible embedding with a given host utterance.

Robustness

Because the tones are spread across all the frames, intentional and unintentional operations such as bit truncation, lowpass filtering, etc. are not likely to cause a loss of data. This may particularly be the case if the tones are close to lower formants, for instance. Some of the encoding techniques of the covert audio message are such that a few errors in the coded data can still reconstruct the audio, albeit with reduced speech quality. It has been found that, for instance, the reconstructed speech from GSM-coded covert utterance with up to 10 percent of bit errors can still convey the message. Coding using Fourier-Bessel expansion also results in understandable message reconstruction with bit errors [7, 8]. The degradation in speech quality in both cases, however, depends on the location of bit errors. Thus, it is possible to convey the covert audio message from the information recovered from the ratio of frame power to tone power in the received stego frames. The extent of loss in embedded information, and hence the received speech quality, need to be studied further for different levels of attacks on the stego signal.

Payload Capacity

In the triad of requirements for efficient steganography, payload capacity appears to be the least satisfied with only one or two bits embedded per frame. Clearly, adding further security to the covert information – such as spread spectrum encoding – cannot be accommodated unless a very long host is used as cover utterance. The low capacity of the tone insertion technique compares unfavorably with methods that employ psychoacoustical masking phenomenon directly [4, 5, 8]; however, the robustness of the technique with flexibility in the choice of tones and the inherent security and ease they offer in data retrieval may more than compensate for the capacity. Additionally, selecting the frequency of the significant tone from frame to frame using a key contributes to increased security somewhat similar to that by spread spectrum encoding. Extending the number of tones to more than four – to eight for embedding 3 bits per frame, for instance – may result in noticeable artifacts in the stego spectrogram as well as in speech quality.

5. CONCLUSION

A method of embedding covert speech in a cover utterance using tones at two and four frequencies has been described. Initial results using clean and noisy host speeches demonstrate the feasibility of the method. The proposed technique satisfies some of the most important criteria for a successful covert communication system, namely imperceptible embedding and accurate data or message recovery. Embedded message payload, however, is low compared to techniques based on the perceptual masking properties of the human auditory system. Further studies are in progress to establish the robustness and applicability of the technique for any general host utterance.

REFERENCES

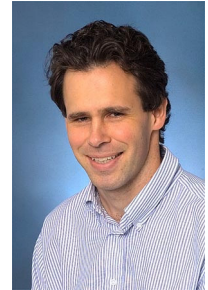
- [1] W. Bender, D. Gruhl, N. Morimoto and A.Lu, "Techniques for data hiding," *IBM Systems Journal*, Vol. 35, Nos. 3 & 4, pp. 313-336, 1996.
- [2] E. Zwicker and H. Fastl, *Psychoacoustics*, Springer-Verlag, Berlin, 1990.
- [3] M.D. Swanson, M. Kobayashi, and A.H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, Vol. 86, pp. 1064-1087, June 1998.
- [4] K. Gopalan, D.S. Benincasa, and S.J. Wenndt, "Data Embedding in Audio Signals," *Proc. of the 2001 IEEE Aerospace Conference*, Big Sky, MT, Mar. 2001.
- [5] K. Gopalan, "Data Embedding in the Phase Spectrum of Speech Signals," *Proc. of the IASTED International Conference on Signal and Image Processing Conference*, Honolulu, HI, Aug. 2001.
- [6] A. Noga, Adjustable Bandwidth Concept Signal Energy Detection Technique, U.S. Patent 5,257,211.
- [7] K. Gopalan, "Speech Coding using Fourier-Bessel Expansion of Speech Signals," *Proc. of the IEEE Industrial Electronics Conference (IECON'01)*, Denver, CO, Nov.-Dec. 2001.
- [8] K. Gopalan, "Audio Steganography for Embedding Compressed Speech," *Proc. of the IASTED Signal and Image Processing Conference*, Kauai, HI, August 2002.

K. 'Gopal' Gopalan (Senior Member, IEEE) received the B.E. degree in Electrical Engineering from P.S.G. College of Technology (University of Madras), Coimbatore, India, in 1971, the M.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Kanpur (IITK), India, in 1974, and the Ph.D. degree in Engineering from the University of Akron, Akron, OH, in 1983. From 1974 to 1979 he was employed at IITK first as Instrumentation Engineer and later as Research Engineer. While pursuing graduate studies at the University of Akron, he held the positions of Teaching Assistant, Research Assistant and Instructor. From 1983 to 1985, he was Assistant Professor of Electrical Engineering at Lafayette College, Easton, PA. Since 1985, he has been with the Department of Engineering at Purdue University Calumet, Hammond, IN, currently holding the position of Professor of Electrical and Computer Engineering. From 1987 to 1995 he conducted research in



the areas of signal and image processing for nondestructive evaluation of advanced materials at Argonne National Laboratory, Argonne, IL, first as a summer faculty research participant and later as a consultant. In addition, he has been a summer faculty research associate at Wright-Patterson Air Force Base, OH (1993), and Rome Laboratory of the Air Force Research Laboratory (1996, 1998, 2000 and 2002), and worked in the areas of speaker identification, analysis of speech under stress, and audio steganography. Dr. Gopalan is the author of the textbook, *Introduction to Digital Microelectronic Circuits* (Chicago: Irwin/McGraw-Hill, 1996).

Stanley J. Wenndt received his B.S. degree with Distinction in Electrical Engineering from Iowa State University in 1987. He received his M.S. and Ph.D. from Colorado State University in 1991 and 1997 respectively. His graduate research focused on speaker identification research in regards to new models and new features to aid in robust algorithms. During this time, he was selected for the Palace Knight program with the U.S. Air Force. Palace Knight is a highly selective training and development program for civilian scientists and engineers. He is currently employed with the Air Force Research Laboratory where his interests are in speaker identification, confidence scores, open-set modeling, dialect identification, speaker count, channel normalization, and noise removal.



Andrew Noga received the B.S. degree in electrical engineering from Clarkson College of Technology, Potsdam NY, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from Syracuse University, Syracuse NY, in 1988 and 1995, respectively. He has been an engineer with the Rome Research Site of AFRL, Rome, NY since 1983 where his work has focused on discrete-time communications signal processing, including detection and parameter estimation, multivariate statistical analyses, spectral characterization, demodulation, and cochannel interference mitigation.

Darren Haddad received a B.S. degree in electrical engineering from the Rochester Institute of Technology, Rochester, NY, in 1991, and a M.S. in electrical engineering from Syracuse University, Syracuse NY, in 1999. He is currently pursuing a Ph.D in electrical engineering at Syracuse University, Syracuse NY. He has been an engineer with the Rome Research Site of AFRL, Rome, NY since 1995 where his work has focused on audio and speech processing, including audio hiding, voice stress analysis, and audio coding.

Scott Adams received a B.S. degree in Electrical and Computer Engineering from Clarkson University, Potsdam, NY in 1987 and an M.S. degree in Computer Engineering from Syracuse University, Syracuse, NY in 1993. From 1987

to the present he has conducted and managed research at the Air Force Research Laboratory, Information Directorate, Multi-Sensor Exploitation Branch, Rome NY, in image processing and exploitation, image registration, and spectral imagery. The current focus of his research is Digital Data Embedding Technologies (steganography, watermarking, steganalysis, and data forensics).

