Spectral Magnitude Speech Steganography

Tamer Rabie
Associate Professor of Computer Engineering
Department of Electrical & Computer Engineering
University of Sharjah, UAE

Driss Guerchi
Associate Professor of Electronics Engineering
School of Engineering, Applied Science and Technology
Canadian University of Dubai, UAE

ABSTRACT

This paper presents a speech-hiding framework that exploits spectral properties of the Fourier magnitude and phase of digital speech signals for the purpose of hiding secret speech messages inside other speech signals for secure transmissions over unsecured networks. The technique used exploits low-pass spectral properties of the speech magnitude spectrum to embed a secret speech signal in the low-amplitude-high-frequency regions of the host speech signal's spectral magnitude. Experimental evaluations on real male and female voice segments show that our technique is capable of hiding one speech message inside another host speech segment to produce a stego speech segment that is indistinguishable from the original host speech, while being able to extract the hidden speech message without any perceived degradations in quality.

General Terms:

Steganography, Multimedia Information Hiding.

Keywords:

Voiced Speech Hiding, LPC, CELP, FFT, Magnitude Spectrum, Phase Separation.

1. INTRODUCTION

The proliferation and exchange of digital voice messages over unsecured networks within social media applications has brought with it issues of how to optimally secure and transmit such large-sized speech files, while retaining the intelligibility of the speech signal.

Information hiding techniques, commonly known as steganography when dealing with hiding secret messages into a cover medium to form a "stego" medium [1], or watermarking when copyright protection of multimedia data is involved [2], have received a great deal of attention in the past decade [3, 4, 5, 6, 7]. Motivated by growing concern about the protection of intellectual property on the Internet and by the threat of a ban for encryption technology, the interest in information hiding systems has been increasing over the years [8].

Techniques for data-hiding inside digital speech signals have been generally confined to four popular schemes, namely; Least Significant Bit (LSB) substitution, Shift Spectrum (SSA), Spread Spectrum (SS) and Frequency Masking (FM). All of them take advantage of the masking property of the Human Auditory System (HAS). In the first case, the secret message is hidden into the least significant bits of the host signal [9]; in the second case, the highest coefficients of the host signal hide the secret message [10]; in the third one, interleaved samples of the host signal are selected to hide the secret message [10]; while in the fourth case, every sample of the secret message is hidden into one sample of the host signal if a masking criterion is satisfied [11].

The highest Hiding Capacity (HC) varies among the above schemes; LSB allows hiding a secret message with the same time-scale of the host signal (HC = 100%), FM allows hiding up to the same time-scale (HC $\leq$ 100%), SS and SSA allows hiding a speech signal up to the half of the host time-scale (HC $\leq$ 50%) [12].

For the purpose of speech message size reduction, speech coding applications have been proposed which aim at hiding bandwidth extension information into narrowband speech signals [13]. After encoding and transmission, this information is used at the decoder-end to reconstruct the wideband signal. These techniques have the advantage of transmitting a wideband signal at the same bit rate as the narrowband signal while conserving the backward compatibility with existing decoders.

It has been shown that the intelligibility of a speech sentence is retained if the inverse transform of the Fourier phase of a long segment of the speech signal is combined with unity magnitude to obtain the phase-only equivalent speech [14]. In fact, in listening to this processed sentence, total intelligibility is retained although the speech has the general quality associated with high-pass filtering and the introduction of additive white noise. The magnitude-only speech has some structure which provides a speech-like characteristic but with no speech intelligibility.

Figure 1 is taken from [14] to show the importance of the phase component spectrum over the magnitude spectrum. Figure 1-(a) is the spectogram of a segment of speech. Figure 1-(b) is the magnitude only spectogram of the same speech segment, and Figure 1-(c) is the phase only spectogram combined with a unity magnitude. It is clear how the phase spectogram highly resembles the original speech spectogram, while the magnitude spectogram bears no resemblance.

In this paper we thus describe a speech-in-speech hiding framework for the purpose of reduced storage and secured transmission requirements of uncompressed RAW speech formats such as the popular wave format, which builds on the Fourier-domain data hiding paradigm, introduced in [15] and applied to preliminary speech hiding in [11].

The rest of this paper is organized as follows. In section 2 we discuss the properties of narrowband speech and present the general framework for our speech hiding scheme. Section 3 the

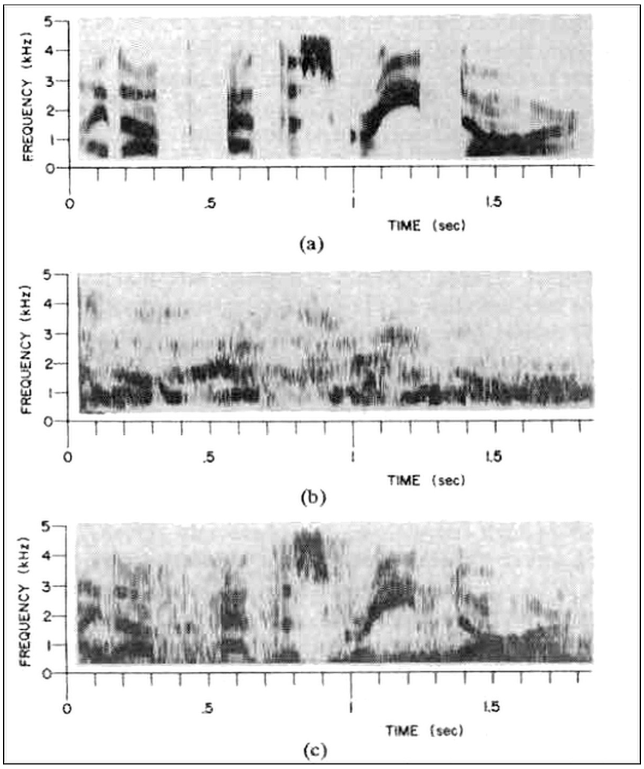

Fig. 1: (a) Spectogram of the original speech segment, (b) magnitude only spectogram of the same speech segment, (c) phase only spectogram combined with a unity magnitude (figure courtesy [14]).

mathematical formulation used in modelling the speech signal to be hidden is presented. Section 4 introduces the main algorithm in detail, and experimental results are evaluated in section 5. Finally concluding remarks are given in section 6.

2. NARROWBAND SPEECH PROPERTIES

Narrowband speech is a low-pass signal where most of the relevant formants are confined to a bandwidth of 4 kHz [16]. In the Code-Excited Linear Prediction (CELP) model, the spectral envelope of a narrowband speech signal is modeled by ten Linear Prediction Coding (LPC) coefficients. These coefficients model the first four formants in the speech signal. In voiced speech, the magnitude spectrum shows very weak components at high frequencies, as is clear from figure 5-(a). Even though unvoiced speech presents larger magnitudes at high frequencies, the intelligibility of the speech signal is negligibly affected if we make some errors in modeling these frequencies. This has motivated us to embed another signal in the low-amplitude-high-frequency part of a host speech signal.

3. CELP MODEL

A secret speech signal, $s_2(n)$,$n = 0, \cdots, 79$ to be hidden in a cover host speech signal $s_1(m)$, $m = 0, \cdots, 159$ is subject to an LP analysis, closed-loop analysis, and a fixed-codebook search in order to extract ten LPC coefficients, pitch parameters, and fixed-codebook contributions, respectively [17]. The speech parameters are updated every 10 ms for a sampling frequency of 8 kHz. The complete algorithm for extracting the CELP model parameters is described in details in [18]. In the following subsections, we review briefly some of the CELP model analysis. It is worth mentioning that no quantization is required in the current application, since none of the CELP model parameters will be transmitted. These parameters will be hidden in the cover host signal $s_1(m)$ in their original unquantized format.

3.1 Linear Prediction Analysis

In the Linear Prediction (LP) analysis, we apply to the signal $s_2(n)$ a 30 ms asymmetric window which consists of two parts: the first part is a half Hamming window and the second part is a quarter of a cosine period. We use a 5 ms lookahead from the future speech. A 10-order predictor is employed on the windowed speech, $s_2^{'}(n)$, to estimate the spectral envelope of the speech signal $s_2(n)$. The predicted signal is given by

$$\hat{s}_2(n) = \sum_{i=1}^{10} a_i s_2^{'}(n-i). \quad (1)$$

The LPC vector $A(z) = (a_1, a_2, \cdots, a_{10})$, used in equation (1), is computed by minimizing the error,

$$e(n) = s_2(n) - \hat{s}_2(n), \quad (2)$$

between the original and the predicted samples. The LPC coefficients are then converted to ten Line Spectral Frequencies (LSF) parameters $w_i, i = 1, \cdots, 10$. Unlike the LPC coefficients, the LSF parameters are all positive since they belong to the normalized frequency domain $[0, \pi]$.

3.2 Pitch Analysis

An open-loop analysis is performed once per 10-ms speech frame to estimate the pitch lag. It is followed by a closed-loop analysis, which refines the search of the pitch delay and pitch gain. The pitch analysis is done once per 5-ms subframe. Since our algorithm is based on 10-ms speech frames for the hidden signals, two pitch delays, T_1 and T_2, and two pitch gains g_1 and g_2 will be added to the hidden information.

3.3 Fixed-Codebook Contribution

The signal after LP analysis and pitch analysis is coded using an algebraic codebook with four pulses per 5-ms subframe. For a 10-ms frame, 8 pulse positions, $P_{1,1}$, $P_{1,2}$, $P_{1,3}$, $P_{1,4}$, $P_{2,1}$, $P_{2,2}$, $P_{2,3}$, $P_{2,4}$, from subframes 1 and 2 are computed. These are added, with their sign indices, $S_{1,1}$, $S_{1,2}$, $S_{1,3}$, $S_{1,4}$, $S_{2,1}$, $S_{2,2}$, $S_{2,3}$, $S_{2,4}$ and two fixed codebook gains, gp_1, gp_2 to the hidden information. At the end of the CELP model analysis, each 10-ms frame of signal $s_2(n)$ will be hidden in terms of its CELP model parameters H_2 in the 20-ms signal $s_1(m)$. Table 1 shows the parameters of the hidden vector H_2.

4. SPEECH-HIDING ALGORITHM

Our speech hiding algorithm is illustrated in figure 2. We start with a 20-millisecond (ms) cover host signal $s_1(m)$, $m = 0, \cdots, 159$ which is transformed to the frequency domain by applying a Fast Fourier Transform (FFT), followed by decomposition into its magnitude and phase spectra, as given in equation (3). A 10-ms hidden message signal $s_2(n)$, $n = 0, \cdots, 79$ is then analyzed using a CELP model, and the extracted CELP parameters, H_2, of

Table 1. : The 32 Parameters of the hidden vector H_2

H_2 Component	Parameters
Line Spectrum Frequencies	$w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}$
Adaptive-codebook Delay	T_1, T_2
Adaptive-codebook Gain	g_1, g_2
Pulse Positions	$P_{1,1}, P_{1,2}, P_{1,3}, P_{1,4}, P_{2,1}, P_{2,2}, P_{2,3}, P_{2,4}$
Pulse Signs	$S_{1,1}, S_{1,2}, S_{1,3}, S_{1,4}, S_{2,1}, S_{2,2}, S_{2,3}, S_{2,4}$
Pulse Gains	gp_1, gp_2

$s_2(n)$ are hidden in the low-amplitude, high-frequency region of the magnitude spectrum of signal $s_1(m)$.

$$S_1(k) = |S_1(k)| \cdot e^{j\varphi(k)}, (k = 0, \cdots, 159) \tag{3}$$

4.1 The Embedding Process

In the first stage of the speech-hiding algorithm, the speech spectrum, $S_1(k)$, of the cover host speech signal $s_1(m)$ is separated into its magnitude spectrum $|S_1(k)|$, $k = 0, \cdots, 159$, and its phase spectrum $\varphi(k)$. The last 32 samples of the first half of $|S_1(k)|$ are replaced by the 32 CELP parameters H_2 of the secret speech signal $s_2(n)$:

$$|S_1(49 : 80)| = H_2(1 : 32). \tag{4}$$

An inverse FFT (iFFT) is performed in order to construct the 20-ms composite (stego) signal $s_3(m)$, $m = 0, \cdots, 159$ as follows:

$$s_3(m) = \text{iFFT}(|S_1| \cdot e^{j\varphi}), (m = 0, \cdots, 159). \tag{5}$$

Signal $s_3(m)$ is the stego signal which contains the 20-ms signal $s_1(m)$ as well as the 10-ms secret speech signal $s_2(n)$ hidden inside it.

4.2 The Extraction Process

The extraction process is illustrated by the block diagram in figure 3. Extracting the 32 CELP parameters from the stego speech signal is conducted in the reverse order to the embedding process; the stego speech signal is transformed to the frequency domain by applying the FFT operation and the magnitude spectrum is separated from the phase spectrum. The 32 CELP parameters are then extracted from the same locations they were embedded in the spectral magnitude of the stego speech signal. These 32 parameters are then used to reconstruct the hidden speech message segment that was embedded. Since the CELP parameter values that are extracted have the exact same values as the embedded CELP, the reconstructed hidden speech signal suffers no perceived degradations in quality.

5. EXPERIMENTAL EVALUATION

We have conducted several simulations to evaluate the performance of our proposed technique, in terms of objective and subjective measures. The evaluation simulations have been conducted on 12 host speech signals uttered by six male and six female speakers and 10 hidden speech signals. Among the 120 possible combinations, we performed 12 simulations by embedding randomly one of the hidden signals in the host signals.

An informal listening comparative test has been performed as a subjective measure. Naive speakers had to listen to both the original host speech signal $s_1(m)$ and the composite (stego) signal $s_3(m)$ in random order and give their preference for the one with better quality. The listeners found it difficult to notice any difference between both signals.

As an objective measure, we select the Segmental Signal-to-Noise Ratio (SNR_{seg}). Instead of working on the whole signal domain, the SNR_{seg} is used to calculate the SNR values of short segments (15 to 20 ms). It is given by:

$$SNR_{seg} = 10log_{10}\left(\frac{\sum_{m=0}^{159}[s_1(m)]^2}{\sum_{m=0}^{159}[s_1(m) - s_3(m)]^2}\right). \tag{6}$$

The SNR_{seg} is most commonly used to measure the quality of reconstruction in a speech waveform by comparing the reconstructed waveform with the original signal. This measure is less sensitive to minor deviations between signals and will be adopted for evaluating our results.

To study the impact of speech hiding on the spectrum of the cover speech, we will also use the weighted spectral distortion (SD_W) measure of Paliwal and Atal [19], where smaller values indicate higher correlation between cover and stego magnitude spectra.

In Table 2, we present the average SNR_{seg} values as well as the average SD_W for the selected simulations. It is clear that speech hiding in female speech cover signals allows for higher performance than using male speech as cover host signals.

Table 2. : Objective Performance of our Scheme

Speaker	Avg SNR_{seg} (dB)	Avg SD_W (dB)
Female	28.94	0.461
Male	26.30	0.507
Average	27.62	0.484

Figure 4 shows an original cover host signal $s_1(m)$ and the stego speech signal $s_3(m)$. Figure 5 shows the original magnitude spectrum of the cover host speech signal $s_1(m)$ and the magnitude spectrum after hiding the 32 CELP parameters of signal $s_2(n)$ in its low-amplitude-high-frequency region. The 32 CELP parameters have replaced the original values in the low-amplitude high-frequency region of the magnitude. These 32 parameters have very small values which is why they appear as zeros in figure 5-(b)

6. CONCLUSIONS

This work has developed a framework for speech-in-speech hiding that can help in reducing the storage and transmission requirements of digital voice messages exchanged over social media applications, as well as for steganography applications of hiding secret speech messages for transmission security over unsecured networks. The technique used exploited the low-pass nature of narrowband speech signals to embed the 32 CELP parameters of another hidden speech signal in the low amplitude high frequency regions of the spectral magnitude of the host

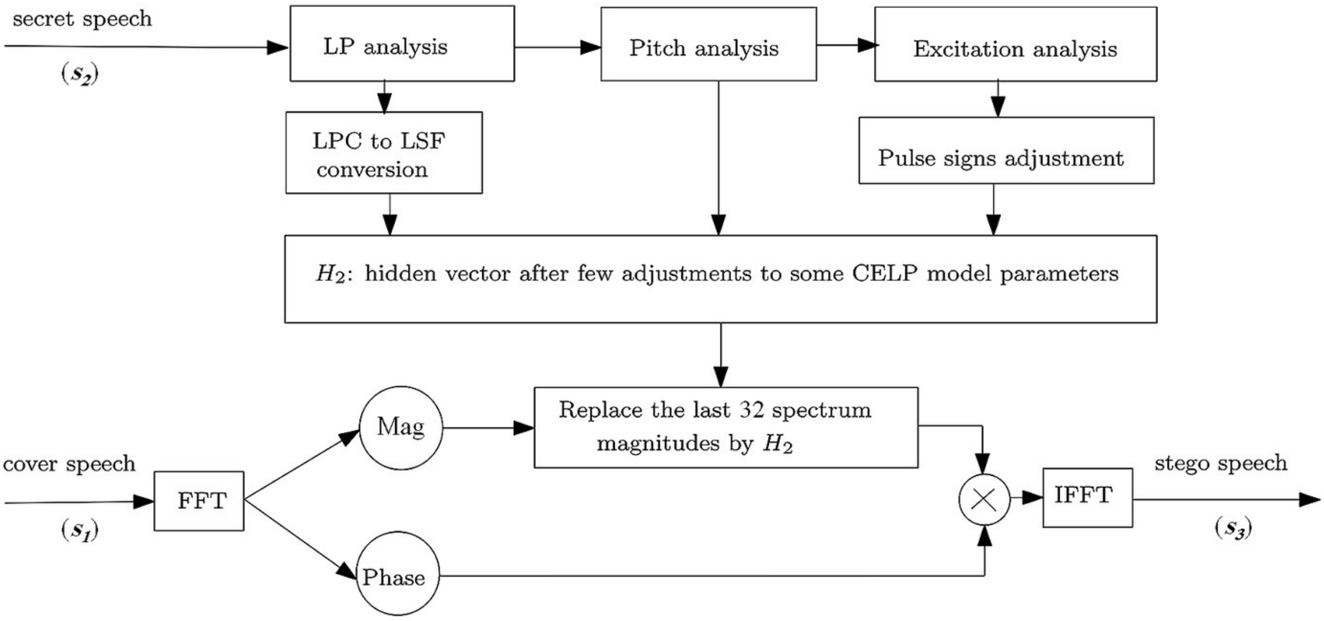

Fig. 2: Block diagram showing the general steps to embed a ***secret speech*** signal (s_2) inside a ***cover speech*** signal (s_1) to produce the composite ***stego speech*** signal (s_3).

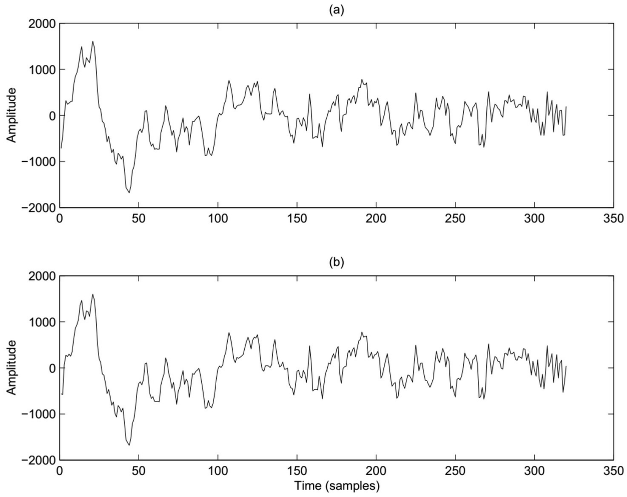

Fig. 4: (a) Original signal $s_1(m)$, (b) Stego signal $s_3(m)$ after hiding the 32 CELP parameters of signal $s_2(n)$. It is clear that the two are almost indistinguishable.

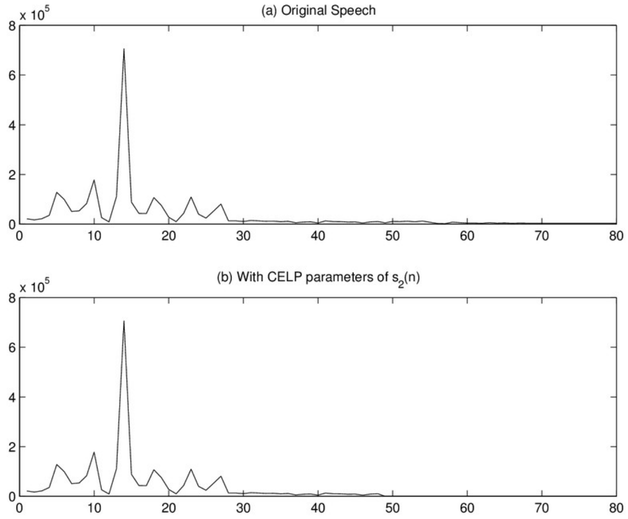

Fig. 5: (a) Magnitude Spectrum of cover host signal $s_1(m)$, (b) Magnitude Spectrum of the stego signal $s_3(m)$. The 32 CELP parameters have replaced the original values in the low-amplitude-high-frequency region of the magnitude. These 32 parameters have very small values which is why they appear as zeros in (b).

speech signal. Experimental results on real male and female voice segments have shown that our technique is capable of hiding one speech message inside another host speech segment to produce a stego speech segment that is indistinguishable from the original host speech, while being able to extract the hidden speech message without perceived degradations in quality. In future work, we would like to extend our paradigm to embedding hidden speech in the commonly used MPEG-Layer-3 (MP3) sound format. The main challenge we would face in this case is how to reduce the effect of compression on the embedded CELP parameters such that the extracted parameters suffer the least amount of degradation which, in turn, would affect the reconstructed speech.

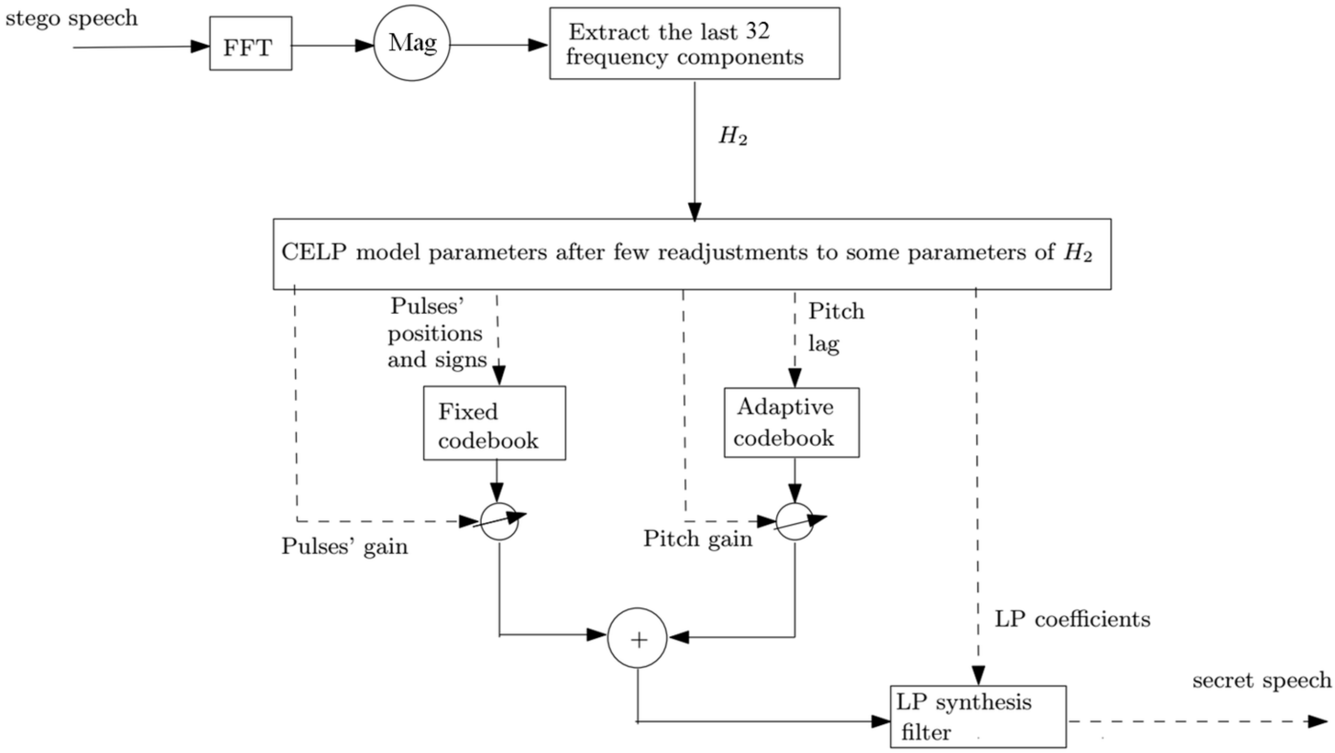
stego speech
FFT
Mag
Extract the last 32 frequency components
H_2
CELP model parameters after few readjustments to some parameters of H_2
Pulses' positions and signs
Pitch lag
Fixed codebook
Adaptive codebook
Pulses' gain
Pitch gain
+
LP coefficients
LP synthesis filter
secret speech

Fig. 3: Block diagram showing the Extraction process to recover the 32 CELP parameters from the same locations of the spectral magnitude of the stego speech signal that they were embedded in.

7. REFERENCES

[1] Niels Provos and Peter Honeyman, "Hide and seek: An introduction to steganography," in *IEEE Security & Privacy Magazine*, pp. 32–44. IEEE Computer Society, 2003.

[2] Min Wu and Bede Liu, "Data hiding in image and video:part I - fundamental issues and solutions," *IEEE Trans. Image Processing*, vol. 12, no. 6, pp. 685–695, June 2003.

[3] Chi Kwong Chan and L.M. Cheng, "Hiding data in images by simple LSB substitution," *Pattern Recognition*, vol. 37, pp. 469–474, 2004.

[4] Kaushal Solanki, Noah Jacobsen, Upamanyu Madhow, B. S. Manjunath, and Shivkumar Chandrasekaran, "Robust image-adaptive data hiding using erasure and error correction," *IEEE Trans. Image Processing*, vol. 13, no. 12, pp. 1627–1639, December 2004.

[5] A.K. Jain, U. Uludag, and R.L. Hsu, "Hiding a face in a fingerprint image," in *Proc. of the International Conference on Pattern Recognition (ICPR)*, Quebec City, Canada, August 2002.

[6] Lisa M. Marvel, Jr. Charles G. Boncelet, and Charles T. Retter, "Spread spectrum image steganography," *IEEE Trans. Image Processing*, vol. 8, no. 8, pp. 1075–1083, August 1999.

[7] Koichi Nozaki, Michiharu Niimi, Richard O. Eason, and Eiji Kawaguchi, "A large capacity steganography using color bmp images," in *ACCV '98: Proceedings of the Third Asian Conference on Computer Vision-Volume I*, London, UK, 1998, pp. 112–119, Springer-Verlag.

[8] Fabien AP Petitcolas, Ross J Anderson, and Markus G Kuhn, "Information hiding-A survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.

[9] Dora M Ballesteros L and Juan M Moreno A, "Real-time, speech-in-speech hiding scheme based on least significant bit substitution and adaptive key," *Computers and Electrical Engineering*, vol. 39, no. 4, pp. 1192–1203, 2013.

[10] Dmitriy E Skopin, Ibrahim MM El-Emary, Rashad J Rasras, and Ruba S Diab, "Advanced algorithms in audio steganography for hiding human speech signal," in *Advanced Computer Control (ICACC), 2010 2nd International Conference on*. IEEE, 2010, vol. 3, pp. 29–32.

[11] Tamer Rabie and Driss Guerchi, "Magnitude spectrum speech hiding," in *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*. IEEE, 2007, pp. 1147–1150.

[12] Dora M Ballesteros L and Juan M Moreno A, "Highly transparent steganography model of speech signals using efficient wavelet masking," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9141–9149, 2012.

[13] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 05)*, Lisbon, Portugal, September 2005, pp. 1497–1500.

[14] A.V. Oppenheim and J.S. Lim, “The importance of phase in signals,” *Proc. of the IEEE.*, vol. 69, no. 5, pp. 529–541, May 1981.

[15] Tamer Rabie, “Frequency-domain data hiding based on the Matryoshka principle,” *Special Issue on Advances in Video Processing and Security Analysis for Multimedia Communications, Int. J. Advanced Media and Communication*, vol. 1, no. 3, 2007.

[16] Douglas O’Shaughnessy, *Speech Communications: Human and Machine*, Wiley-IEEE Press, 2 edition edition, November 1999.

[17] D Guerchi, H Harmain, T Rabie, and E Mohamed, “Speech secrecy: An fft-based approach,” *International Journal of Mathematics and Computer Science*, vol. 3, no. 2, pp. 1–19, 2008.

[18] R. Salami, C. Laflamme, J.-P Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, “Design and description of cs-acelp: A toll quality 8 kb/s speech coder,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 116–130, March 1998.

[19] Kuldip K Paliwal and Bishnu S Atal, “Efficient vector quantization of lpc parameters at 24 bits/frame,” *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 1, pp. 3–14, 1993.