Synthetic speech detection and audio steganography in VoIP scenarios*

Daniele Capolupo[1] and Fabrizio d'Amore[2]

[1] University of Tor Vergata, Rome
Enterprise Engineering Dpt.
danielecapolupo@outlook.it
[2] Sapienza University of Rome
Dpt. of Computer, Control, and Management Engineering
damore@dis.uniroma1.it

Abstract. The distinction between synthetic and human voice uses the techniques of the current biometric voice recognition systems, which prevent that a person's voice, no matter if with good or bad intentions, can be confused with someone else's. Steganography gives the possibility to hide in a file without a particular value (usually audio, video or image files) a hidden message in such a way as to not rise suspicion to any external observer. This article suggests two methods, applicable in a VoIP hypothetical scenario, which allow us to distinguish a synthetic speech from a human voice, and to insert within the Comfort Noise a text message generated in the pauses of a voice conversation. The first method takes up the studies already carried out for the Modulation Features related to the temporal analysis of the speech signals, while the second one proposes a technique that derives from the Direct Sequence Spread Spectrum, which consists in distributing the signal energy to hide on a wider band transmission.
Due to space limits, this paper is only an extended abstract. The full version will contain further details on our research.

Keywords: synthetic detection, modulation, temporal feature, steganography, information hiding, speech signal covert communication, data embedding, spread spectrum, signal processing

1 Introduction

This paper was developed starting from the hypothesis that automatic systems generate conversations using a synthetic voice, whose breaks, where a comfort noise is usually inserted, contain text messages, to be sent to a specific recipient through a VoIP phone call. The idea of the scenario is to use the artifice of the synthetic voice as a tool to confuse a hypothetical observer, thus giving the

* This work has been partially supported by the TENACE PRIN Project (n. 20103P34XC) funded by the Italian Ministry of Education, University and Research.

possibility to transmit hidden messages from one point to another without the risk of being caught. In this scenario there are two important phases: first, the recognition of a synthetic voice from the human one, taking as object of analysis a VoIP conversation intercepted and rebuilt; second, the introduction of the steganographic method, which could be used in VoIp conversations, but in our case it will be applied directly to audio files of conversations.

In [2] a biometric voice verification system was initially proposed. The development of such systems implies very deep studies for safety purposes, to prevent that false identities are exchanged for real ones [9, 13]. To prevent a false identity to be mistaken for a real one, the current safety studies start with the discrimination of a human voice from a synthetic one. This work uses techniques of vocal signals and aims at illustrating two new algorithms (MelFCC and MagM) able to recognize human voices. Our research was inspired by the apporoach in [12], where techniques for voice verification/conversion were introduced. Usually, in these techniques the voice signal is not input to directly the system that performs the processing of the voice, but it is first transformed into a more compact and meaningful representation [12], that allows to obtain a set of properties named "features" [10], that use the well-know Mel scale[8], which approximates the human auditory system response. The Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up the Mel-Frequency Cepstrum (MFC), which are a short-term representation of power spectrum of a voice signal. A cepstrum is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal and the power cepstrum in particular finds applications in the analysis of human speech[1].The MFCCs (Mel Frequency Cepstral Coefficients[1, 8]) are calculated for each single frame without knowing the next frame (frame by frame), therefore, it is very difficult to capture the correlations between frames, or the temporal characteristics. On the other hand, the frame-based operation in the process of synthesizing voice can introduce temporal artifacts [11]. In order to consider the subsequent frame dependence and therefore capture temporal artifacts generated by synthetic voice, modulation properties were used to develop another algorithm for extraction of features, together with the implementation of an additional technique that is able to discriminate the human voice from synthetic one, named Magnitude Modulation. The modulation features, derived from the amplitude spectrum, carry a long term temporal voice information and therefore, are able to detect temporal artifacts due to the frame by frame processing in synthesizing voice signal [3, 4, 7]. The modulation features were used to capture the audio frame change in order to recognize the synthesized voice. In our study, to carry out discrimination, were produced two identical versions (length and vocabulary used) by two different voice conversations, one containing the human voice, the other with a synthesized voice. Each conversation was divided into parts of equal length and each of them has been subjected to the analysis of algorithms implemented. Similar partitioning was performed on individual conversations after have been transmitted through a VoIP phone call, whose client used G. 711 codec.

In many VoIP telephony services, a signal called Comfort Noise (CN) is introduced into the conversation breaks. The presence of this signal allows partners to have the feeling that the line remains on, even during long or above average breaks. This signal is typically generated directly to the terminals, but, at the source side, you can insert a hidden signal that contains a text message M, which, as such, is recognized as a network communication activities and therefore encoded as if it were a voice, tricking monitoring system, located downstream from the recipient, mistaking for CN. Since the direct insertion of a text message, appropriately converted to a digital signal within the CN, would produce a peak clearly visible with a simple frequency signal analysis, this work proposes a steganographic technique, known in signal-processing as Direct Sequence Spread Spectrum (DSSS). The DSSS allows to distribute the same earlier information (energy of the converted text message) on a wider transmission bandwidth, thus eliminating the peak emerged from the frequency analysis of the background noise. This process does not change the overall power of the signal, which must be sufficiently higher than noise to allow reconstruction at the nodes, because the decrease in average spectral power density (dBm/Hz) is compensated by the enlargement of the bandwidth [5, 6].

2 Synthetic Voice Detection

Some methods to recognize the synthetic voice from the human voice work at frame level, i.e. splitting the signal into segments and lead, in the first instance, separate analysis on them. Some of these techniques are focused on the study of the characteristics of the signal amplitude such as the MFCC (Mel Frequency Cepstral Coefficient), others, such as the MGDCC (Modified Group Delay Cepstral Coefficients), exploit some properties of the phase of the signal (and, therefore, its group delay). Both MFCC and MGDCC do not allow to derive correlations between different frames as well as the temporal characteristics of the extracted features. MM methods (Magnitude Modulation) and PM (Phase Modulation) are respectively based on previous MFCC and MGDCC, but they introduce additional processing that make possible to engage the temporal relations between different frames, projecting features in an analysis of medium-long term. This work used exclusively structured analytic methods of signal amplitude, i.e. the MFCC and MM.

2.1 MFCC (Mel-Frequency Cepstral Coefficients)

To derive the MFCC coefficients of a given signal it is necessary to process it by the analysis of Short Time Fourier Transform (STFT), assuming that the signal is quasi-stationary within a short period (for example a 25 ms window).

The STFT of a signal voice $x(n)$ is as follows:

$$X(\omega) = |X(\omega)|e^{i\Phi(\omega)} = X_{real}(\omega) + iX_{imaginary}(\omega)$$

where $|X(\omega)|$ is the magnitude spectrum, $\Phi(\omega)$ is the phase spectrum and $i = \sqrt{-1}$. We note that $X(\omega)$ has two parts: real part

$$X_{real}(\omega) = \sqrt{(X_{real}(\omega))^2 + (X_{imaginary}(\omega))^2}$$

and imaginary part

$$X_{imaginary}(\omega) = \arctan(\frac{X_{imaginary}(\omega)}{X_{real}(\omega)})$$

The power spectrum is defined to be $|X(\omega)|^2$.

Before running the STFT, and then derive the MFCCs, we must preprocess the speech signal and divide it into separate windows. The pre-processing consists in:

1. The speech signal is divided into overlapping segments of equal size with duration of 25 ms, called frames, with a percentage of overlap between consecutive frames of 50%.
2. Each frame is multiplied by a "window function," in our case we used the Hamming window function needed to mitigate the effect that would create, in the subsequent extraction of features, if you used a finite-size segment, thinning the edges which lie at the beginning and end of each frame and avoiding ghostly artifacts.

After terminated the pre-processing, the Mel-Frequency Cepstral Coefficients are obtained for each frame window using the following steps:

1. Apply the FFT (Fast Fourier Transform) to compute for each frame the spectrum $X(\omega)$ of $x(n)$.
2. Compute the power spectrum $(|X(\omega)|)^2$.
3. Process the Filter-Bank Energies (FBE) applying the Mel frequency filter-bank to the power spectrum $(|X(\omega)|)^2$.
4. Apply the Discrete Cosine Transform (DCT) to access the scale of the FBE and enable us so to calculate the MFCCs.

Features extraction from MFCC

1. The signal is decomposed into frames consisting of n points with m shift points.
2. For each frame the above analysis described is conduct to obtain c Mel coefficients.
3. From the c Mel coefficients, new coefficients are derived, defined by the two following functions "DELTA" and "DELTA-DELTA":
   DELTA. Consider a vector x of K elements

$$x(0), x(1), \ldots, x(K-1)$$

The DELTA is the function $\delta_{x,N}$ defined as follows:

$$\delta_{x,N}(k) = \begin{cases} x(k) & k \in \{-1, \ldots, N\} \cup \{K-N, \ldots, K\} \\ \frac{\sum_{i=1}^{N} i(x(k+i)-x(k-i))}{2\sum_{i=1}^{N} i^2} & k \in \{N, N+1, \ldots, K-N\} \end{cases}$$

DELTA-DELTA. Once defined $\delta_{x,N}$, the DELTA-DELTA is defined as follows:

$$\delta\delta_{x,N}(k) = \begin{cases} \delta_{x,N}(x(k)) & k \in \{-1, \ldots, N\} \cup \{K-N, \ldots, K\} \\ \frac{\sum_{i=1}^{N} i(\delta_{x,N}(x(k+i))-\delta_{x,N}(x(k-i)))}{2\sum_{i=1}^{N} i^2} & k \in \{N, N+1, \ldots, K-N\} \end{cases}$$

Notice that $\delta\delta_{x,N}(k) = \delta_{x,N}(\delta_{x,N}(k))$. The coefficients computed through DELTA and DELTA-DELTA are known as the "differential" and the "acceleration" coefficients, respectively. Because the voice seems to contain dynamic information that is distributed over time and because the MFCC feature vectors only describe the power spectral envelope of a single frame, the DELTA coefficients allow to calculate the trajectories of the MFCC coefficient over time. The DELTA-DELTA coefficients are calculated in a similar way, but this time starting from DELTA and not from the MFCC static coefficients.

4. For each frame is obtained a vector of length $3c$.

If N is the number of frames where the signal has been split, connecting the N vectors of length $3c$, we get a super vector of length $3cN$. The variance of this super vector is the feature of the signal under test.

2.2 Magnitude Modulation (MM)

The modulation features, obtained from amplitude spectrum, contain information on long-term temporal features of voice signal, and therefore are able to detect temporal artifacts due to the frame by frame processing in synthesizing voice signal. The modulation features were thus used to capture frame variations for synthesized voice.

Features extraction from Magnitude Modulation (MM)

1. Divides the signal into frames, consisting of n_{FFT} points with m_{FFT} shift samples.
2. Defines a "segment" consisting of n consecutive frames of the type described in Section 1, with m frame shift. The value of n must be sufficiently large to allow the capture of temporal information.
3. From the spectrum of each frame gets c coefficients of the filtered banks by Mel scale (c Mel-scale filter-bank coefficients). It defines "path filter-bank" the n points set of the same coefficient of Mel.
4. Applies a MVN (Unitary Mean Zero Variance) for each trajectory of filter-bank to normalize the average and variance from zero to one.

5. For each normalized trajectory, calculates the power density spectrum, calculated with a different number of FFT points (nS) (considering only the positive frequencies and therefore consists of $\frac{nS}{2}$ points).
6. For the current window, gets an array that consists of c rows and $\frac{nS}{2}$ columns, vector given by concatenating of the c rows and $\frac{nS}{2}$ coefficients.
7. Due to the large size and to the high correlation of the spectral modulation, derived from different trajectories of filter-bank, it is necessary to reduce the size by applying a Principal Component Analysis (PCA) method. For each window, it is calculated the maximum PCA variance among the possible c (the PCA is applied considering as variables the Mel's coefficients and the frames as observations).

At the end of the above described process, if K is the number of windows where the signal has been devided, we get K variances. In this experimental work we used as feature of the signal the average (arithmetic mean) of the K maximum variances.

2.3 Experiments and results

The study covers the extraction of the features related to the conversation files (syntetic and human), which were divided in 17 and 16 parts. The features to extract are 33 for the audio signal with human voice and 33 for the audio signal with synthetic voice. Each of the $2N$ files available ($N = 33$ human and $N = 33$ synthetic) were submitted to MelFCC (Mel Frequency Cepstral Coefficients) and MagM (Modulation Magnitude) analysis. To evaluate the performance of the two methods, the study adopts the EER (Error Equal Rate). The EER is the error rate that results when the percentage of the human voice is incorrectly classified as a synthetic one and it is equal to the percentage of synthetic voice incorrectly classified as a human voice. The two parameters considered were the FAR (False Acceptance Rate) and FRR (False Rejection Rate). The FAR, specifies how often the system is tricked, i.e. when a synthetic voice gets wrongly perceived as a human voice. The FRR, specifies the frequency with which the system fails to indicate that a human voice is truly human. Lower is the EER value, better will be the performance. In order to calculate the variance of the trajectories produced by the models between the different signals (synthetic and human) there was produced a simulation with the two methods of analysis proposed (MelFCC and MagM).

Each point on the blue curve corresponds to a feature of one of the 33 parts of 2 human conversation files. Similarly, each point of the green curve is given from the feature derived from the 33 parts of 2 synthetic conversation files. The curves shown in Figure 1 are the result of an experimental standardization process, whose steps will be described in detail. This work is based on the assumption that events logging of digital files, can be characterized, from a statistical point of view, in such a way that the features associated with human speech, as well as those relating to synthesized conversations, are determinations of a variable of a stationary and ergodic random process.

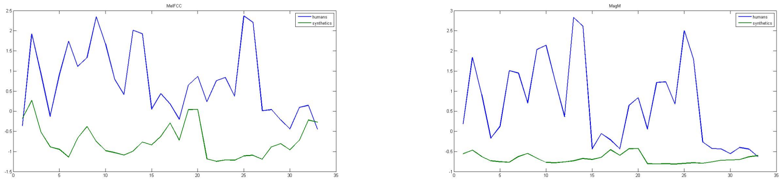

Fig. 1. MelFCC and MagM graphs curves

We consider three random variables: r, which is extracted from the $R(t)$ process, associated with the "real signal" E_r, with expected value m_r; variable s, that is extracted from process $S(t)$ concerning the event "synthetic signal" E_s, with expected value m_s; variable x, extracted from the process $X(t)$, associated with the event "signal" E_x given by the union event "real signal" with the event "synthetic signal" ($E_x = E_r \cup E_s$). We adopt the following assumptions:

1. Variable r has N determinations in a R set of equally probable values.
2. Variable s has N determinations in a set S of equally probable values.
3. Variable x has $2N$ determinations in a set X of equally probable values (as of R and S), where the first N determinations belong to set R and the determinations by the $(N+1)$-th to the $2N$-th belong to set S.
4. Both $R(t)$ and $S(t)$ are stationary and ergodic (then $X(t)$ is the same).

The expected value of the random variable x is

$$E(x) = \sum_{i=1}^{2N} \frac{1}{2N} x_i = \sum_{i=1}^{N} \frac{x_i}{2N} + \sum_{i=N+1}^{2N} \frac{x_i}{2N} = \frac{m_r}{2} + \frac{m_s}{2} = m_x$$

that is given by the arithmetic mean of the two expected values m_r and m_s.

Thus, proceeding to a experimental normalization, as depicted in Figure 2, if x is the vector of the $2N$ determinations of the random cumulative variable x, taking away m_x to each of $2N$ determinations (or features), it gets a new random variable with expected value zero. In order to make it with single variance, simply compute the variance of the new random variable $(x - m_x)$ and split each of the new $2N$ for this determination. As shown in Figure 2, the straight line parallel to x-axis through the origin is a little "watershed" between the features of human signals and those of the synthetic conversations.

Symmetrical thresholds vectors have been considered in respect to the origin, namely vectors th with N_t components, whose values belong to $\{-M, -M+1, \ldots, M-1, M\}$, attributing to the first cell th[0] the score 0% and to last cell th$[N_t - 1]$ the score 100%. The values used in the simulations were $M = 3$ and $N_t = 240$. The false positive (False Acceptance) happens when a determination (that is a feature of a synthetic signal) is greater than the threshold. Similarly the false negative FR (False Rejection) happens when a value of r (in other words, a feature of a real signal) is less than the threshold.

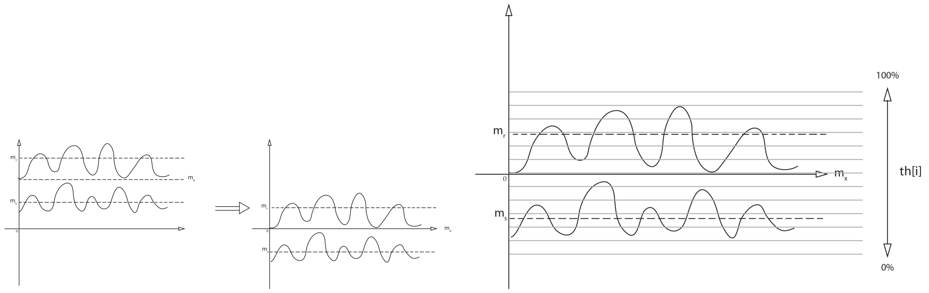

Fig. 2.

We denote by Fr the vector of false negative FRR (False Rejection Rate), with N_t length, and by Fa the vector of false positives FAR (False Acceptance Rate). For each threshold value th[i], Fr(th[i]) is the number of points on the curve of real signals that are under the threshold, divided by N, while the Fa(th[i]) is the number of points on the curve synthetic signals which are above the threshold, divided by N. The Fr curve increases with the value of th[i], while Fa decreases. There will be a th[j] threshold value for which Fr(th[j]) = Fr(th[j]) = EER (Equal Error Rate).

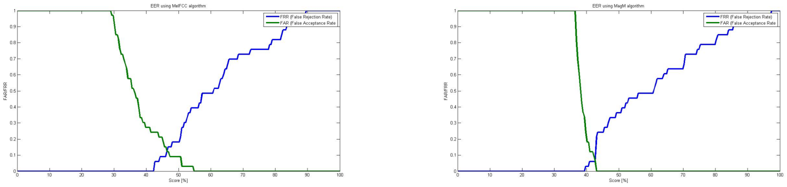

Fig. 3. MelFCC and MagM EER graphs.

2.4 VoIP conversations analysis with MelFCC and MagM algorithms

Also in this case the study proceeded by applying the algorithms MelFCC and MagM to human voice files and to synthesized voice files, obtained through a VoIP connection. The following figures report the obtained results and show there is not a substantial difference in the operation of the algorithms we used. EER curves for both algorithms demonstrate that it is possible to discriminate a human voice from a synthetic one, even for VoIP conversations, confirming the greater goodness (more value the EER) of MagM algorithm than the MelFCC. In Table 1 we can observe EER values obtained from the simulations.

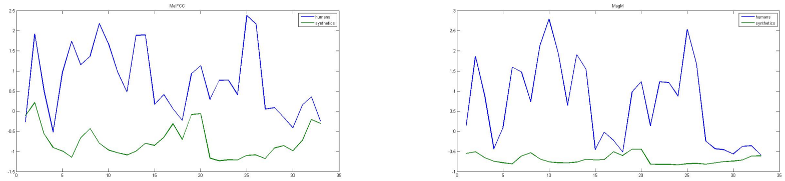

Fig. 4. MelFCC and MagM VoIP graphs curves.

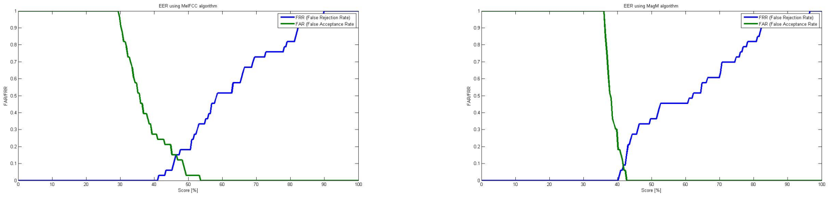

Fig. 5. MelFCC and MagM EER VoIP graphs

3 Steganography: Direct Sequence Spread Spectrum

In Figure 6 we have a representation, both in time and frequency domain, of a 50 ms duration Comfort Noise signal, sampled at 32 KHz and estimated spectrum of 4096 FFT points. Supposing to have a bits sequence, derived from a text encoding, as a message M to be transmitted. Each character is encoded with n_B bits (for example 8 in ASCII). A message of a length of n_C characters, will produce a $n_C n_B$ bits length message to be transmitted. Consider for example these possible initial 18 bits, obtained by encoding some text characters: 01111000, 11010111, and 11. T_b is the bit time. With reference to the sampling rate of 32

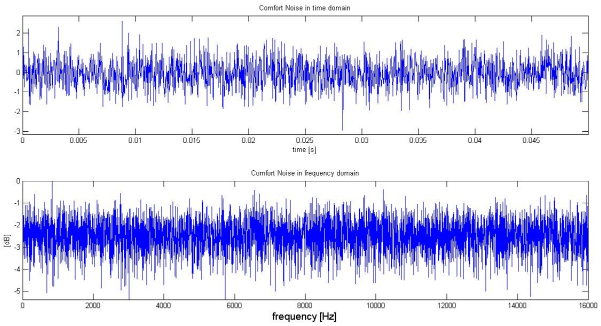

Fig. 6.

Table 1. EER Values

	No-VoIP	VoIP
MelFCC	0.1212	0.1515
MagM	0.0606	0.0758

kHz, it follows that a single bit corresponds to $n_{\text{samplePerBit}} = T_b f_s$ samples. Setting the interval bit to 32 ms and having sampling rate like 32 kHz we get a $n_{\text{samplePerBit}}$ of 1024 samples. Therefore, the previous 18-bit sequence will be used in a window of a 576 ms time and composed of 18432 samples. A signal of this type will never be transmitted because its drastic changes in amplitude will produce very large high frequency components, thus requiring a very large transmission bandwidth. To avoid this issue, the signal can be filtered with a low-pass filter with appropriate cut-off frequency (approximately $\frac{0.5}{T_b}$) and translated in audio band frequency, centered in a frequency carrier of $f_c = \frac{f_s}{4} = 8$ KHz.

We can now insert this signal inside the Comfort Noise, then a 3dB Signal to Noise Ratio, the signal strength to be inserted will be about twice that of the comfort noise. Looking at Figure 7, we can realize that in the domain of time it is pretty hard to note an anomaly, but in the frequency domain we notice a visible peak frequency.

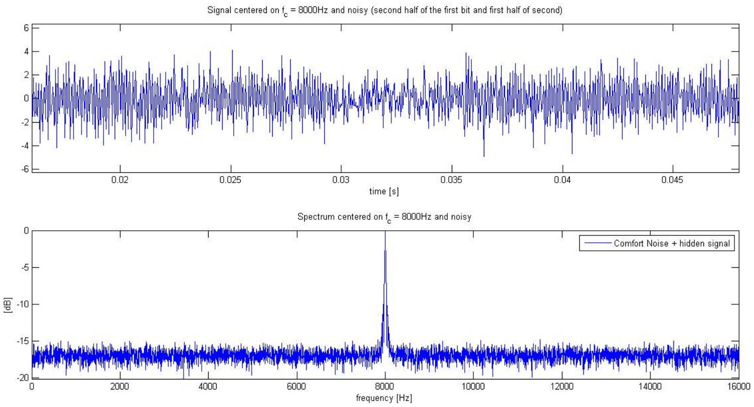

Fig. 7.

In order to improve this method, and inspired by papers [5, 6], we introduce a technique called Spread Spectrum, which allows to eliminate the peak that previously emerged of about 16 dB from the background noise. For example, Figure 8 shows two signals with associated power equal to about -44 dBm (40nW), but having respectively (in blue), power physics average spectral density, of about

$\frac{-71\ dBm}{\text{Hz}}$ $(\frac{79.5\ pW}{\text{Hz}})$, for a band of 500 Hz, and that wider (green) power physics average spectral density of about $\frac{-86\ dBm}{\text{Hz}}$ $(\frac{2.5\ pW}{\text{Hz}})$, for a 16 KHz band.

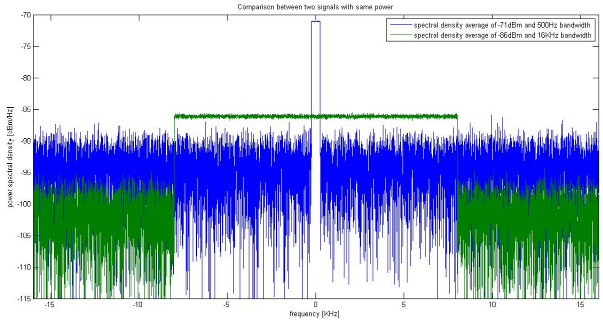

Fig. 8.

Figure 8 shows that if we decrease the power spectral density, we can get results by widening the bandwidth, thus, if the signal carries a binary message, also a considerable increase in throughput, i.e. reduction of bit time T_b. In order to broaden the signal bandwidth, without increasing the bit transmission speed (therefore decreasing T_b) we must multiply the signal, each bit time T_b, for another signal, named Spreading, that for simplicity is always the same for each bit interval (but in theory it could be used an algorithm to modify it) consisting of a number S_f (Spreading factor) of bits not carrying any information, because they are repeated periodically, each of which has a chip time durability $T_c = T_b/S_f$. It follows that the number of samples for each interval (T_c) chips will be a S_f-th for each bit interval (T_b). According to the general principle presented, we decided to multiply the signal, each bit time $T_b = 32$ ms, for a Pseudo Noise sequence of spreading, to ensure a more evenly distributed spectrum in the existence band. Using a spreading sequences of this type, spectrum will be as flat as possible. Selecting a spreading factor $S_f = 256$, the chip time of PN sequence (which will multiply the signal bit time for bit time) will be $T_c = \frac{T_b}{S_f}$, i.e. $32 \cdot 10^{-3}/256 = 125\ \mu s$.

Therefore we have again a difficult distinction in the time domain, but this time the distinction will be difficult even in the frequency domain, since there will no longer be the presence of a carrier wave with a peak spectral density that emerges from the depths of about 16 dB. In Figure 9 we can observe a comparison between the first and the second technique of steganography.

The last comparison locates the only critical point of the technique previously presented. In the first technique we did a simple analysis of the signal density spectrum to detect an anomaly, but the Spread Spectrum technique can be noticed just by considering the signal power density spectrum to the second power. Indeed, although it is difficult to see any trace of the modulated signal in

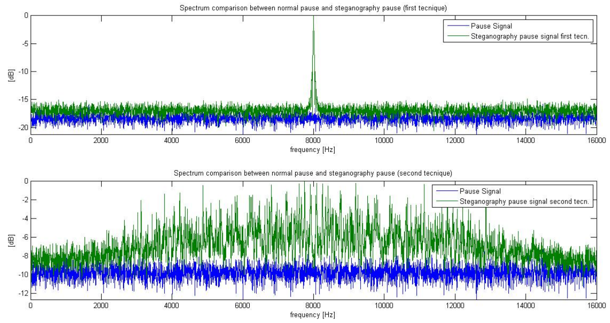

Fig. 9.

occult way, if we look at the spectrum in Figure 10, we can notice the presence of a spectral density peak at $f = f_c = 8$ KHz and two other equally spaced peaks from f_c.

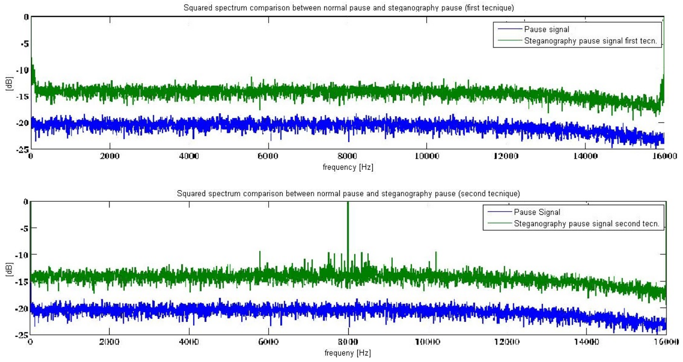

Fig. 10.

3.1 Application and results

To implement the steganography algorithm, the 33 synthetic files were examined to identify the files with long pauses. From the examination 3 files were chosen, and then the study applied the steganography algorithm, getting the same results for each one. The text message that was transmitted and subsequently recovered was “Rome22May,” indicating a place and a day of the week, to simulate the will of sending occult logistic and temporal informations.

1. Pre-processing phase

Used variables:

- Hiddentext = “Rome22May” 9 characters
- Number of samples for bit $n_{\text{SampleForBit}} = 136$
- Frequency sampling $f_s = 22050$ Hz
- Bit time T_b, $n_{\text{SampleForBit}}/f_s = 0.0062$ s
- Number of samples for chip $n_{\text{SampleForChip}} = 2$

After identifying the time period containing the longest pause and inserted the comfort noise signal within all pauses, the hiding message has been converted to bits by using the 8 bit ASCII encoding, for a total of 9 characters of 8 bits, namely 72 bits. The bit sequence derived from ASCII conversion was used to generate the digital signal using BPSK modulation, rectangular filter and oversampling factor equal to $n_{\text{samplePerBit}}$.

2. Spread Spectrum phase

Used variables:

- SpreadingFactor S_f, $\frac{n_{\text{SampleForBit}}}{n_{\text{SampleForChip}}} = 68$
- Chip time T_c, $T_b/S_f = 9.12 \cdot 10^{-5}$

Once modulated, the spreading binary sequence must have good spectral properties, looking like a white noise. For this reason we used the MATLAB function `randint`, and BPSK modulation with oversampling factor of $\frac{n_{\text{SamplePerBit}}}{S_f} = 2$. The generated spreading sequence was multiplied with the previously generated digital signal, for each bit time. In order to reduce the bandwidth occupation, a 256-th order minimum phase filter (FIR) was applied, low-pass with $\frac{f_c}{2} = 2$ cut-off frequency, where $f_c = \frac{1}{T_c} = 2$, T_c = chip time equal to $\frac{T_b}{S_f}$. The spreading signal was filtered and used to modulate a carrier frequency f_p and summed to the comfort noise signal to be inserted into the pause of the audio file selected (see Figure 11).

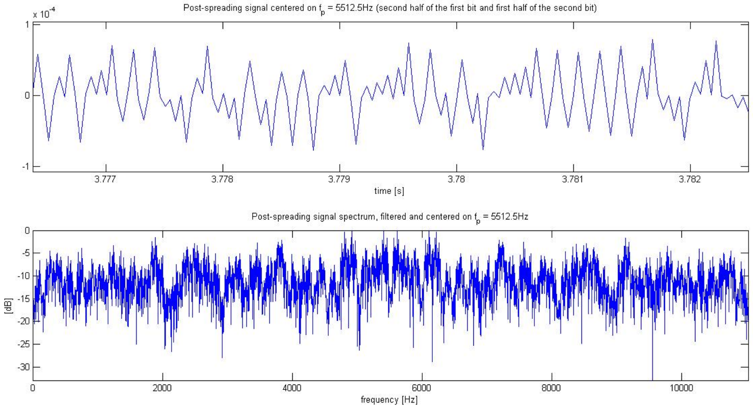

Fig. 11.

3. De-Spreading phase

   The De-Spreading phase requires that the recipient is aware of the following parameters:

   - startSample and stopSample of the signal audio received.
   - $n_{\text{SamplePerChip}}$ value and SpreadingFactor value.
   - f_p carrier frequency where the signal was modulated.
   - Spreading/De-Spreading sequence, composed by S_f binary symbols.

   From the audio signal received, containing the hidden text message, the portion in question was extracted for the process of steganography, and restored in the base-band. This is obtained by multiplying the signal by $\cos(2\pi f_p t)$ and applying a lowpass filter to remove the $2f_p$ replication. The BPSK signal with a binary frequency equal to $f_c = f_b S_f$, has been demodulated exploiting the small oversampling factor equal to 2, to get a signal with an antipodal levels signal 1, and -1. The reconstruction of the spreading binary signal was obtained by applying the sign function $(1 - \text{sign}())/2$, which passes from the antipodal levels 1 and -1 to the logic levels 0 and 1. The bits sequence that we got, equal to 9792 bit $(N_b \times n_{\text{SamplePerBit}})$, was organized in an array of N_b rows by $n_{\text{SamplePerBit}}$ columns. Each row corresponds to one bit of the ASCII encoded text message which has been sent. To get the despreading bit sequence we carried a series of mathematical operations, applying the XOR function, between the received bits and the spreading bits sequence. The generic r line of the despreading matrix, is given by $\text{XOR}(c, s)$, where c is a binary string of $n_{\text{SamplePerBit}}$ bit, all equals the ASCII encoding bits of the case, and s is the binary version of spreading sequence. Because the receiver knows s, she will compute $\text{XOR}(\text{XOR}(c, s), s) = \text{XOR}(c, \text{XOR}(s, s)) = \text{XOR}(c, z) = c$, where $z = 00 \cdots 0$.

   After the despreading process, for each bit of the binary ASCII sequence of the transmitted text, $n_{\text{SamplePerBit}}$ bits were obtained, which should be all equal to each other and equal to the original text bits transmitted. In the case that something was received incorrectly, the error can be recovered as long as at least $n_{\text{SamplePerBit}}/2 + 1$ are properly received. To recover the 72 bits of the ASCII encoding text transmitted we proceeded to conduct a review with a majority decision for every 136-bit sequence ($\frac{9792}{136} = 72$), which is equal to the value we used for $n_{\text{SamplePerBit}}$.

   The process of recovering the transmitted text was completed by applying the function that transforms the binary sequence obtained into 8-bit ASCII characters, from which was derived the information “Rome22May,” which was been transmitted in covert mode.

4 Conclusions

This work showed the feasibility of the initial scenario assumptions, concerning the transmission using steganography techniques of hidden messages through VoIP calls and the use of a synthetic voice audio signals. In that regard, the Mel

Frequency Cepstral Coefficient (MelFCC) algorithm and the Magnitude Modulation (MagM) algorithms have been developed, based on the analysis of the voice detection techniques proposed in the literature, to discriminate a human voice audio signal from a synthetic voice audio signal. The results showed that these algorithms have basically the same behavior, whether applied to the original file, synthetic or human, whether transmitted through a VoIP phone call. The steganography algorithm proposed, based on the application of the technique called "Spread Spectrum," has allowed us to confirm that we can hide a text within the comfort noise introduced in the pauses (as normally happens in a telephone call) of an audio file.

References

1. Bogert, B.P., Healy, M.J.R., Tukey, J.W.: The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. Proc. Symp. Time Series Analysis, Ed.: M. Rosenblatt, John Wiley pp. 209–243 (1963)
2. Campbell, J.P., J.: Speaker recognition: a tutorial. Proceedings of the IEEE 85(9), 1437–1462 (Sep 1997)
3. Kingsbury, B., Morgan, N., Greenberg, S.: Robust speech recognition using the modulation spectrogram. Speech Communication 25(1-3), 117–132 (1998)
4. Kinnunen, T., Lee, K.A., Li, H.: Dimension reduction of the modulation spectrogram for speaker verification. In: Odyssey. p. 30. ISCA (2008)
5. Nugraha, R.: Implementation of direct sequence spread spectrum steganography on audio data. In: Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. pp. 1–6 (July 2011)
6. Rupanshi, Preeti, V.: Audio steganography by direct sequence spread spectrum. In: International Journal of Computer Trends and Technology (IJCTT),Published by Seventh Sense Research Group. vol. 13, pp. 83–86 (July 2014)
7. Sam, S., Xiao, X., Besacier, L., Castelli, E., Li, H., Siong, C.E.: Speech modulation features for robust nonnative speech accent detection. In: INTERSPEECH. pp. 2417–2420. ISCA (2011)
8. Stevens, S.S., Volkmann, J., Newman, E.B.: A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America 8(3), 185–190 (1937)
9. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. Speech and Audio Processing, IEEE Transactions on 6(2), 131–142 (Mar 1998)
10. Wolf, J.J.: Efficient acoustic parameters for speaker recognition. The Journal of the Acoustical Society of America 51(6B), 2044–2056 (1972)
11. Wu, Z., Siong, C.E., Li, H.: Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: INTERSPEECH. pp. 1700–1703. ISCA (2012)
12. Wu, Z., Xiao, X., Chng, E., Li, H.: Synthetic speech detection using temporal modulation feature. In: ICASSP. pp. 7234–7238. IEEE (2013)
13. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. Audio, Speech, and Language Processing, IEEE Transactions on 17(1), 66–83 (Jan 2009)