Security of Digital Files: Audio Tampering Detection

Sebastian-Alexandru ARGHIRESCU
Faculty of Electronics, Telecommunications and Information Technology,
National University of Science and Technology POLITEHNICA Bucharest, Romania
sarghirescu@upb.ro

Abstract

In an age where most of the data we interact with daily is stored digitally, methods of checking its authenticity become more and more essential. This is especially true for sound, as the increasing public availability of AI models makes tampering with audio files easier than ever. In this paper, we will be investigating the current landscape of audio forensics as well as our new hardware-based solution for double encoding detection.

Index terms: audio, automation, security, spectrum, tampering

1. Introduction and state of the art

There are two main types of approaches in classical audio tampering detection: active approaches based on embedding a watermark inside the file during or after recording and passive approaches which do not require any additional embedding.

A passive approach for determining if an audio file has been tampered with is by doing an Electric Network Frequency (ENF) analysis on the recording in question. This is based on the finding that we can extract the mains hum from the recording and compare it with a database to detect tampering. The mains hum fundamental frequency should be at 50 Hz or 60 Hz depending on the location where the audio recording was taken. Due to the loading of national electric grind in various countries, the harmonics of the mains hum also differ by city and by time of day, so it is possible to also get an estimate of the time of the recording [1]. One instance of such ENF analysis is presented in [2] where the authors use FFT (Fast Fourier Transform) and machine learning algorithms to classify extracted ENF signals from signals of different lengths. The detection accuracy found varied by the algorithm used, signal duration (5, 15, 25, 35, 45 s) and type of tampering (copy, deletion) but ultimately fell in the 81-99% detection range. Another approach for ENF analysis is presented in [3] in which the authors use DFT (Discrete Fourier Transform) and a parallel RDTCN-CNN (Residual Dense Temporal Convolutional Network - Convolutional Neural Network) approach to obtain higher accuracy at 97-98%. As shown in [3], the main disadvantage of such methods is that they are highly dependent on the choice of training data set as well as requiring adequate hardware for the models to run on. A model that works in a country with over 90% accuracy could fall to below 60% in another country and would have to be tested and retrained every time on deployment.

Looking at active approaches, these are usually based on some form of embedding a DWT (Discrete Wavelet Transform) watermark into the source recording which can then be used to detect tampering or to reconstruct the original signal from various integrity fractions. One such example is [4] where the authors use a compressed version of the original audio signal to generate the watermark. This method showed a higher signal-to-noise ratio for the watermark than other active approaches, over 93% tampering detection rate and 80-98% recovery of the original recording with a destruction rate between 50-20%. The main disadvantage of this method is that it is in itself altering the original

signal. If a passive approach is used on a watermarked recording, then it will output a false positive. In this case, we would need to know that a watermark was previously applied to the recording for an accurate tampering detection. The accuracy of active approaches also does not seem to be superior to that of their passive counterparts described above.

Another important subject in audio tampering detection is that of double encoding. This concept is less relevant in voice recording, where the classical methods discussed above are used, and more relevant in music. We will be measuring the compression level of a musical recording using Equation 1, where F_S is the sampling frequency of the recording and N_{bps} is the number of bits per sample.

$$F_S * N_{bps} = \text{Bitrate} \quad (1)$$

Looking at the example in Figure 1, we see the spectrogram for the same song (Adventure Found Me by Jason Graves) four times at different bitrates: 128, 192, 256, and 320 kbps encoded using LAME [5] in a .MP3 container. The spectrograms were drawn using Spek [6].

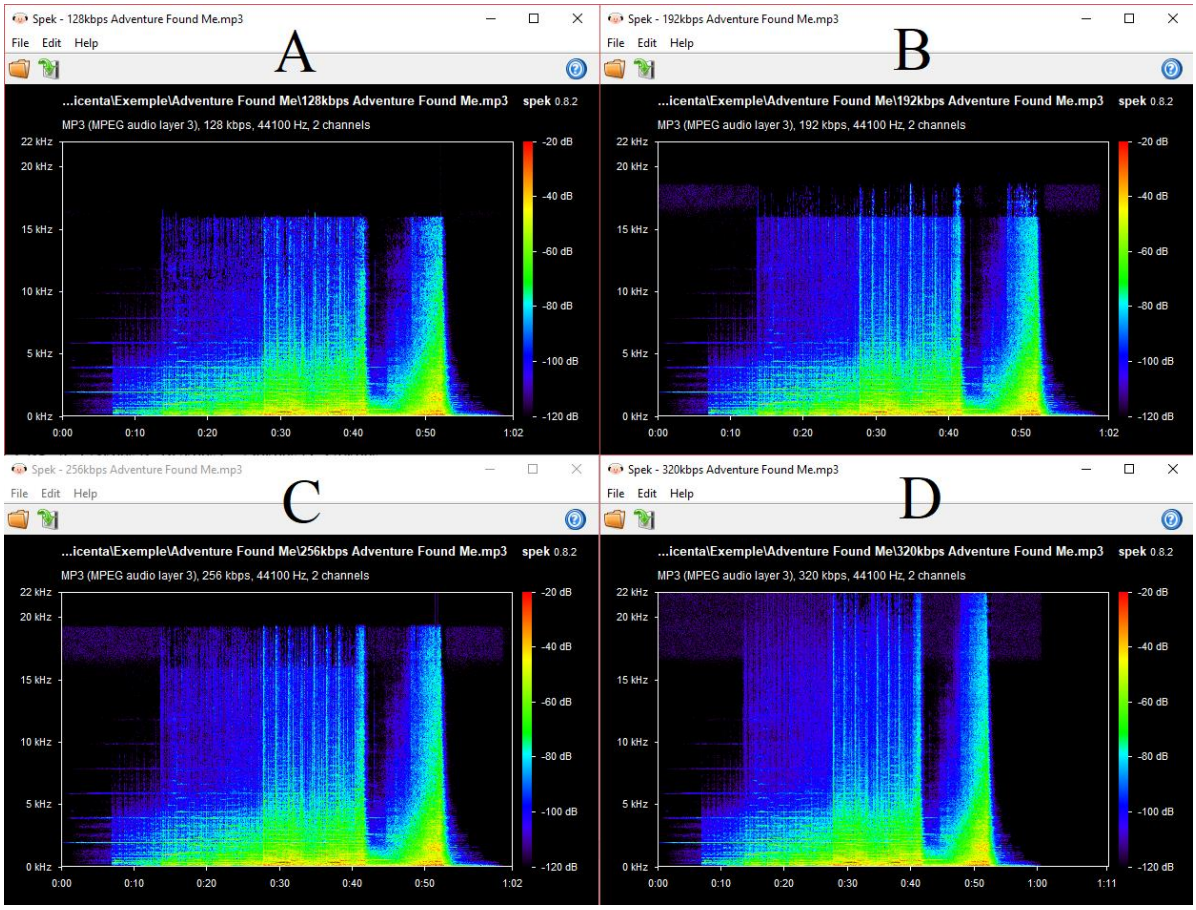

Fig. 1. Spectrogram for the song Adventure Found Me - Jason Graves, 44.1 kHz, MP3, CBR, A. 128 kbps, B. 192 kbps, C. 256 kbps, D. 320 kbps

It was observed in [7] that we can use a spectrogram to determine the compression level of an audio file. We can see that for each jump in compression, the high-frequency content of the recording is visibly cut. This is normal as it reduces the size of the file, but the 128 kbps file could be re-encoded at 320 kbps increasing back the size without regaining any of the lost information. This is called double-encoding. In Figure 2 we can see that the double-encoded file on the left is almost the same size as the original file on the right, but its compression level is much higher. We can also see that the metadata of the file is the same which means the only way to distinguish the two is by looking at their spectrum. It is worth noting that depending on the file and the compression algorithm used the double encoded file may be larger than the original. The file size and reported bitrates in Spek and in Windows File Explorer are marked in red.

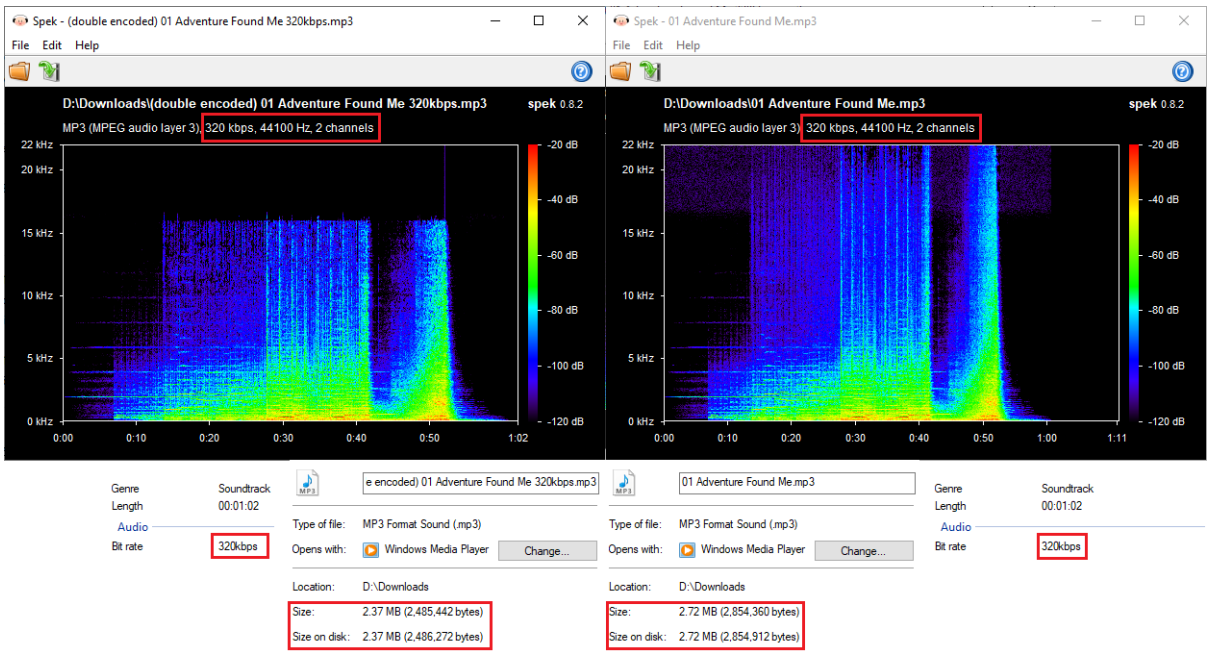

Fig. 2. File properties and spectrogram of the double encoded recording of Adventure Found Me - Jason Graves (left) and the original version (right)

Looking at the state of the art for this type of audio tampering we find that most solutions are using various machine learning algorithms. One example of this is [8] in which the authors use transformer networks to look at the compression history of audio files. The detection accuracy they obtained was between 84 and 94%. Another example of this is [9] which is more focused on codec classification rather than tampering detection but is using similar methods. Even though some of these methods such as [8] are less susceptible to the dataset used for training a big disadvantage of them is that they cannot be used to look at encrypted data streams such as the music used by streaming services. Other solutions for double encoding detection that do not use machine learning techniques usually use Modified Discrete Cosine Transform (MDCT) coefficients. Example of this are [10] and [11]. The accuracy of these methods in seems to be higher but the datasets used for testing were smaller and they seem to be very sensitive to the codec used for testing, working best with MP3 files. There is also the manual looking-at-the-spectrogram approach we have presented above but it is very slow and very dependent on the tester's experience.

2. Materials and methods

Our proposed solution uses a development board with additional signal acquisition circuitry in order to calculate the FFT and detect the presence of a specific predetermined audio band in the recording. In case of the double-encoding detection this interest band would be [17-20] kHz. A block diagram for the board cand be found in Figure 3. The red connections in Figure 3 represent power lines while the black connections represent data lines. An Arduino equipped with an ATmega2560 was used as a microcontroller for testing due to the larger memory requirement for the display buffer. The board was powered from a 9V battery in order to avoid as much interference as possible on the audio signal acquisition path. A 3.5mm male-to-male cable was used to inject the signal into the board form a signal generator. The same cable was used when testing the audio file from Figure 1, but instead of a signal generator, it was connected to the audio output of a computer using Windows 10.

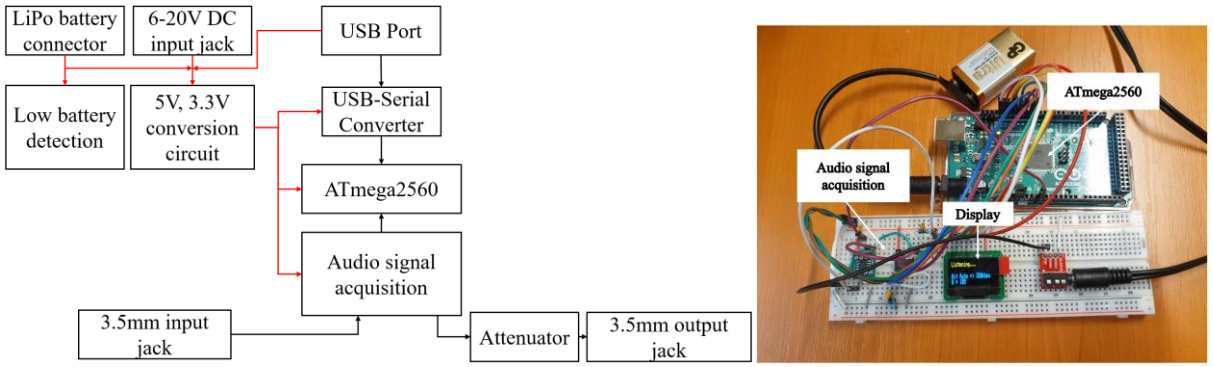

Fig. 3. Board diagram for the development board (left) and a picture of the board prototype (right)

The board samples the input signal at a frequency of 76.9 kHz, storing each sample in an 8-bit variable. The board then calculates the FFT and filters out all the frequency data that is not in the [17-20] kHz interval. It then calculates the spectral power distribution on smaller bands of 1 kHz inside the interest interval and estimates a quality factor Q. Based on this Q, the original bitrate of the song is estimated. The formulas used for these estimations are based on experimental observations. In order to test the performance of our solution we have used a signal generator to generate sine, square and triangle waves at different frequencies and amplitudes. The result of this testing can be seen in Figure 4 below.

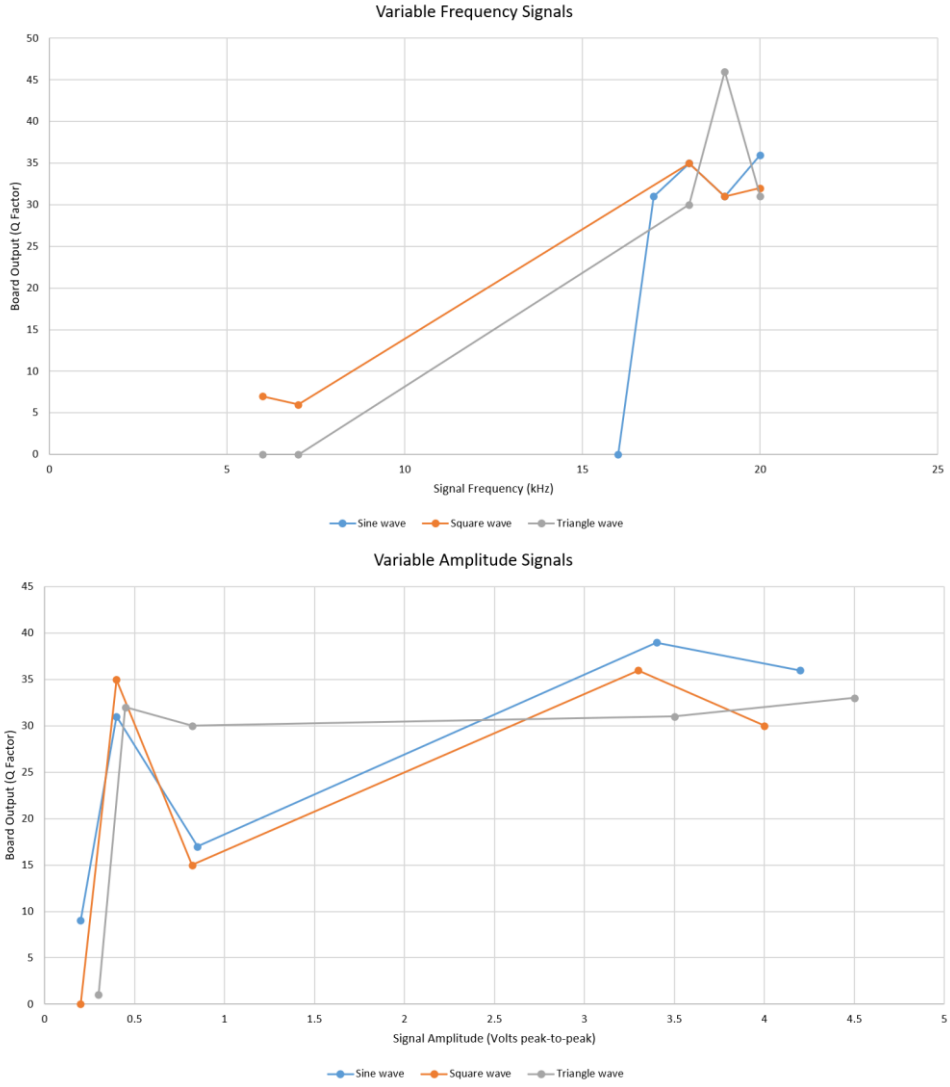

Fig. 4. Board output Q factor on the y-axis and signal frequency (up) or signal amplitude (down) on the x-axis for sine, square and triangle waves used for testing

If we consider the double encoding correctly detected for a Q-factor of 20 or over, then we can calculate a detection accuracy of 1 for the variable frequency signals and 0.8 for the variable amplitude signals for a mean accuracy of 0.9 or 90% for this very small dataset. We have also tested the songs present in Figure 1.A and Figure 1.D and the results can be seen in Figure 5 below.

Fig. 5. Detection results on the board display after the song in Figure 1.A was played (left) and after the song in Figure 1.D (right)

We see that the detection was far more difficult for signals with low amplitudes. We also have to take into account that square and triangle waves have harmonics on multiples of the fundamental frequency, and thus these harmonics could be detected causing false positives. This was not the case here probably because of the relatively low amplitudes of those harmonics in the test signals but could happen on larger datasets. The song we tested in Figure 1 was also not detected correctly, being recognized as 256 kbps instead of 320 kbps. This is likely caused by the short duration of the high frequency bursts observed in Figure 1.

3. Conclusion

There are many ways to tamper with audio recordings, from classical methods such as deleting or moving sections, to more advanced techniques like encoding lower-quality music in larger containers. There are also plenty of methods for detecting such tampering attempts. From active approaches consisting of embedding watermarks in the recording to passive ENF analysis to various machine learning algorithms trained on multiple data sets. Some approaches such as ENF analysis or watermark embedding have high accuracy rates but are unable to detect double encoding. Machine learning algorithms are also accurate but are highly dependent on the dataset used for training and require dedicated complex hardware to host the models. Other solutions using MDCT coefficients can detect encoding-based tampering but are not codec-agnostic.

Our proposed solution has lower detection rates even on pure monotonal signals but it looks at the recording in the analog domain so it can analyze encrypted data streams such as those found on streaming services in close-to-real-time. It is susceptible to low-duration pulses of high-frequency content but is less affected by the codec used in the recording than other methods. The amplitude of the signal must also be high enough (over 0.2 Vpp) for the detection of double encoding to be possible. Real-time ENF analysis may also be possible with this approach, but a lot more research on larger datasets is required.

References

[1]. R. Garg, A. L. Varna and M. Wu, "Modeling and analysis of Electric Network Frequency signal for timestamp verification," 2012 IEEE International Workshop on Information Forensics and Security (WIFS), Costa Adeje, Spain, 2012, pp. 67-72, doi: 10.1109/WIFS.2012.6412627.

[2]. Hsu, H.-P.; Jiang, Z.-R.; Li, L.-Y.; Tsai, T.-C.; Hung, C.-H.; Chang, S.-C.; Wang, S.-S.; Fang, S.-H. Detection of Audio Tampering Based on Electric Network Frequency Signal. Sensors 2023, 23, 7029. https://doi.org/10.3390/s23167029.

[3]. Zeng, C.; Kong, S.; Wang, Z.; Li, K.; Zhao, Y. Digital Audio Tampering Detection Based on Deep Temporal-Spatial Features of Electrical Network Frequency. Information 2023, 14, 253. https://doi.org/10.3390/info14050253.

[4]. Hu, Y., Lu, W., Ma, M. et al. A semi fragile watermarking algorithm based on compressed sensing applied for audio tampering detection and recovery. Multimed Tools Appl 81, 17729-17746 (2022). https://doi.org/10.1007/s11042-022-12719-0.

[5]. LAME MP3 Encoder, available online: https://lame.sourceforge.io/, last accessed on 30.10.2024.

[6]. Spek, available online: https://github.com/alexkay/spek, last accessed on 30.10.2024.

[7]. Alessandro, Brian & Shi, Y.Q. (2009). Mp3 bit rate quality detection through frequency spectrum analysis. 10.1145/1597817.1597828.

[8]. Z. Xiang, P. Bestagini, S. Tubaro and E. J. Delp, "Forensic Analysis and Localization of Multiply Compressed MP3 Audio Using Transformers," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 2929-2933, doi: 10.1109/ICASSP43922.2022.9747639.

[9]. Atieh Khodadadi, Soheila Molaei, Mehdi Teimouri, Hadi Zare, Classification of audio codecs with variable bit-rates using deep-learning methods, Digital Signal Processing, Volume 110, 2021, 102952, ISSN 1051-2004, https://doi.org/10.1016/j.dsp.2020.102952.

[10]. Yang, R., Shi, Y.-Q., & Huang, J. (2009). Defeating fake-quality MP3. Proceedings of the 11th ACM Workshop on Multimedia and Security - MM&Sec ’09. doi:10.1145/1597817.1597838.

[11]. Bianchi, T., Rosa, A.D., Fontani, M. et al. Detection and localization of double compression in MP3 audio tracks. EURASIP J. on Info. Security 2014, 10 (2014). doi:10.1186/1687-417X-2014-10.