Forensic Analysis and Localization of Multiply Compressed MP3 Audio Using Transformers

Ziyue Xiang[1] Paolo Bestagini[2] Stefano Tubaro[2] Edward J. Delp[1]

[1]Video and Image Processing Lab (VIPER), School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA

[2]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

PAPER ID: 4716

icassp 2022 Singapore

Introduction

- MP3 is one of the most popular audio compression methods
- MP3 compression leaves distinct compression artifacts in the compressed bitstream, which can be used for forensics analysis
- Our proposed method can localize parts of an MP3 audio that are multiply compressed—this can be used to detect audio splicing attacks or speech synthesis attacks
- In a typical speech synthesis attack, the attacker decodes an MP3 audio acquired from the Internet into the time domain and replaces a temporal segment with an uncompressed audio signal generated with speech synthesis methods; the spliced audio is compressed again with MP3

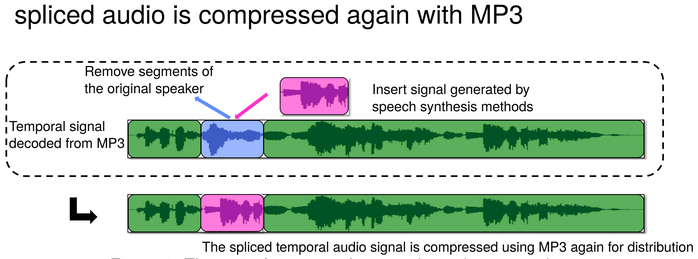

Figure 1: The typical process of a speech synthesis attack.

In the compressed spliced audio, the green region (decoded from MP3) is compressed twice using MP3, whereas the magenta region (generated by speech synthesis methods) is compressed once; our proposed method can localize the compression inconsistency and determine the temporal location of the spliced audio segments

Approach

- Our method uses low-level MP3 encoding parameters to make decisions

Fields	Description
`part_23_length`	Size of coded binary data
`scalefactor`, `scalefac_compress`, `scalefac_scale`, `preflag`	Scalefactor value info
`global_gain`, `subblock_gain`, `big_values`, `region_count`	Quantization step sizes
`table_select`, `count1_table`	Huffman table selection info
`block_type`, `mixed_block_flag`	Sub-band window selection info
`mdct_coef`	Decoded MDCT coefficients

Table 1: The list of MP3 encoding parameters used by our method.

- For each temporal segment (i.e., frame) of the audio signal, we extract MP3 encoding parameters from the compressed signal and generate a corresponding feature vector using different preprocesssing techniques
- Our method analyzes L = 20 frames at a time; the feature vectors from the L frames are processed by a transformer neural network to generate L binary valued labels indicating whether each frame has been multiply compressed or not

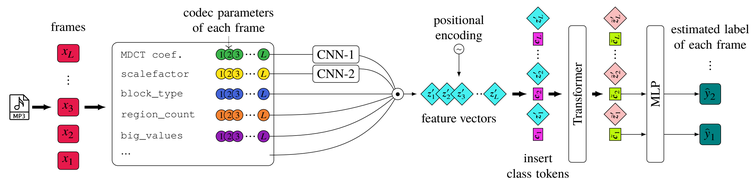

Figure 2: The block diagram of our proposed method.

Results and Conclusion

- We trained and tested our method using uncompressed audio signals from LJSpeech [1], GTZAN [2], and MAESTRO [3]
- We compared the performance of our method against other approaches that used MP3 codec information for multiple compression detection

Method	Jaccard Score	F_1-score	Balanced Accuracy
Yan *et al.* [4]	13.53	18.71	53.35
Yang *et al.* [5]	30.73	40.95	55.28
Liu *et al.* [6]	48.18	58.91	68.72
Our Approach	80.50	84.43	84.49

Table 2: Performance metrics comparison.

	Num. MP3 Compression			
Method	Single	Double	Triple	Overall
Yang *et al.* [5]	67.28	43.51	43.02	43.27
Yan *et al.* [4]	91.71	14.86	15.10	14.98
Liu *et al.* [6]	79.36	57.27	58.85	58.06
Our Approach	84.61	83.76	84.92	84.34

Table 3: The recall of each method against the number of MP3 compressions.

	Last MP3 Comression Type							
Method	C64	C128	C160	C192	V1	V2	V4	V6
Yan *et al.* [4]	17.30	6.86	6.80	4.87	22.80	23.23	22.59	17.55
Yang *et al.* [5]	19.27	70.75	71.71	66.21	45.58	39.05	27.12	22.06
Liu *et al.* [6]	58.45	56.48	54.47	55.01	55.15	56.41	57.47	67.93
Our Approach	73.09	88.06	89.65	90.84	93.31	91.21	83.52	65.51

Table 4: The recall of multiple compression localization for each method against selected last MP3 compression types. CBR compression is denoted by C<bit rate>; VBR compression is denoted by V<quality index>.

- Our proposed method temporally localizes multiple compressions at the frame level, which provides finer localization granularity
- The experiments showed that our method had the best performance compared to other approaches and was robust against many MP3 encoding compression settings
- In the future, we will extend our technique to audio compression methods such as AAC

[1] K. Ito and L. Johnson, *The lj speech dataset*, 2017. [Online]. Available: `https://keithito.com/LJ-Speech-Dataset/`.
[2] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002. DOI: `10.1109/TSA.2002.800560`.
[3] C. Hawthorne, A. Stasyuk, A. Roberts, *et al.*, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," *Proceedings of the International Conference on Learning Representations*, 2019.
[4] D. Yan, R. Wang, J. Zhou, C. Jin, and Z. Wang, "Compression history detection for MP3 audio," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 2, pp. 662–675, 2018. DOI: `10.3837/tiis.2018.02.007`.
[5] R. Yang, Y. Q. Shi, and J. Huang, "Detecting double compression of audio signal," *Media Forensics and Security II*, vol. 7541, pp. 200 –209, 2010. DOI: `10.1117/12.838695`.
[6] Q. Liu, A. H. Sung, and M. Qiao, "Detection of double MP3 compression," *Cognitive Computation*, vol. 2, no. 4, pp. 291–296, 2010. DOI: `10.1007/s12559-010-9045-4`.

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu.