BMC Research Notes

DATA NOTE Open Access

Dataset for file fragment classification of audio file formats

Atieh Khodadadi and Mehdi Teimouri*

Abstract

Objectives: File fragment classification of audio file formats is a topic of interest in network forensics. There are a few publicly available datasets of files with audio formats. Therewith, there is no public dataset for file fragments of audio file formats. So, a big research challenge in file fragment classification of audio file formats is to compare the performance of the developed methods over the same datasets.

Data description: In this study, we present a dataset that contains file fragments of 20 audio file formats: AMR, AMR-WB, AAC, AIFF, CVSD, FLAC, GSM-FR, iLBC, Microsoft ADPCM, MP3, PCM, WMA, A-Law, μ-Law, G.726, G.729, Microsoft GSM, OGG Vorbis, OPUS, and SPEEX. Corresponding to each format, the dataset contains the file fragments of audio files with different compression settings. For each pair of file format and compression setting, 210 file fragments are provided. Totally, the dataset contains 20,160 file fragments.

Keywords: Audio file formats, Classification, File formats, File fragments

Objective

A considerable amount of Internet traffic is used for exchanging audio file formats. As the sizes of these files are usually much bigger than the maximum network packet size, the files are segmented into fragments. The fragments generated by various users are transmitted over the network. Some of these fragments can be received by the network surveillance unit. The network surveillance unit may wish to detect the file format of each fragment for network forensics purposes.

Some researches have been carried in the field of file fragment classification of audio file formats [1–4]. There are a few publicly available datasets of files with different formats [5–7]. Therewith, there is no public dataset for file fragments of audio file formats. This makes it difficult for other researchers to compare the proposed methods with the existing methods.

In this study, we present a dataset that contains file fragments of 20 audio file formats: Adaptive Multi-Rate (AMR), Adaptive Multi-Rate Wideband (AMR-WB), Advanced Audio Coding (AAC), Audio Interchange File Format (AIFF), Continuously Variable Slope Delta modulation (CVSD), Free Lossless Audio Codec (FLAC), Global System for Mobile Communications Full Rate (GSM-FR), Internet Low Bitrate Codec (iLBC), Microsoft Adaptive Differential Pulse Code Modulation (ADPCM), MPEG Audio Layer-3 (MP3), Pulse-Code Modulation (PCM); Windows Media Audio (WMA), A-Law, μ-Law, G.726, G.729, Microsoft GSM, OGG Vorbis, OPUS, and SPEEX. Corresponding to each format, the dataset contains the file fragments of audio files with different compression settings.

Data description

First, the whole set of the uncoded (raw) dataset of speech files is taken from www.voxforge.org [8]. These raw files are then converted in order to obtain audio files in 20 different formats: AMR, AMR-WB, AAC, AIFF, CVSD, FLAC, GSM-FR, iLBC, Microsoft ADPCM, MP3,

*Correspondence: mehditeimouri@ut.ac.ir
Information Theory and Coding Laboratory, University of Tehran, Tehran, Iran

PCM, WMA, A-Law, µ-Law, G.726, G.729, Microsoft GSM, OGG Vorbis, OPUS, and SPEEX. For each audio file format, different compression settings are considered. The raw data for all compression settings of a specific format is the same. However, there is no overlap between the raw data used for different formats.

96 pairs of file format and compression setting are considered. For each pair of file format and compression setting, we have 210 compressed audios. So, totally we have 20,160 audio files. Each of these files is segmented into 1 Kbyte (i.e. 1024 bytes) fragments. Then, one fragment is randomly selected among the fragments of each file. Before randomly selecting the fragments, 12.5% of the initial fragments and 12.5% of the final fragments of each file are discarded. This is to ensure that the fragments do not contain the file headers or trailers.

For each pair of file format and compression setting, we have 210 file fragments. So, the dataset of file fragments contains 20,160 file fragments. The dataset is partitioned according to 20 different file formats. Each partition is represented by an individual data file shown in Table 1. For example, data file 1 (i.e. aac.zip) contains 7 sub data files: aac-8.dat, aac-16.dat, aac-32.dat, aac-48.dat, aac-64.dat, aac-80.dat, and aac-96.dat. Sub data files are provided in a generic binary data file format with .dat file extension. Each sub data file contains 210 fragments.

Data file 21 (i.e. SettingsTable.pdf) contains a table that specifies 96 pairs of file format and compression setting. In this table, the software program employed for generating each file format is also specified. Data file 22 (i.e. ConversionSettings.zip) contains several screenshots of the software programs that display the employed compression settings. Data file 23 (i.e. ReadFragments.m) is a script in MATLAB language that reads all the fragments from one or more sub data files. By running this script and selecting some sub data files, the fragments contained in these sub data files are read and stored in a variable name Dataset. Variable Dataset is a MATLAB cell array with two rows. Each column in this cell array corresponds to one of the selected sub data files. The first element of each column is a string value that specifies the sub data file name. The second element of each column is a structure array with only one field named fragments. Dataset {2, i}(j).fragments (j = 1,2,…,210) is a cell array with length one that contains one fragment of the jth file in the selected sub data file i.

Table 1 Overview of data files/data files

Label	Name of data file/data file	File types (file extension)	Data repository (DOI)
Data file 1	aac	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 2	adpcm	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 3	aiff	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 4	alaw	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 5	amr	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 6	awb	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 7	cvsd	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 8	flac	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 9	g726	Archive file format (.zip) containing one generic binary data (.dat) file	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 10	g729	Archive file format (.zip) containing one generic binary data (.dat) file	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 11	gsm	Archive file format (.zip) containing one generic binary data (.dat) file	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 12	gsmwav	Archive file format (.zip) containing one generic binary data (.dat) file	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 13	ilbc	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 14	mp3	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 15	ogg	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 16	opus	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 17	pcm	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 18	speex	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 19	ulaw	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 20	wma	Archive file format (.zip) containing multiple generic binary data (.dat) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 21	SettingsTable	Portable document format (.pdf)	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 22	ConversionSettings	Archive file format (.zip) containing 97 portable network graphics (.png) files	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)
Data file 23	ReadFragments	Matlab script file (.m)	OSF (https://doi.org/10.17605/OSF.IO/AHCYU)

Limitations

- The size of the fragments is considered to be fixed and equal to 1024 bytes.
- A defined subset of file formats and compression settings are considered.

Abbreviations

AMR: adaptive multi-rate; AMR-WB: adaptive multi-rate wideband; AAC: advanced audio coding; AIFF: audio interchange file format; CVSD: continuously variable slope delta modulation; FLAC: free lossless audio codec; GSM-FR: Global System for Mobile Communications Full-Rate; iLBC: internet low bitrate codec; ADPCM: adaptive differential pulse code modulation; MP3: MPEG audio layer-3; PCM: pulse-code modulation; WMA: windows media audio.

Acknowledgements

Not applicable.

Authors' contributions

MT designed the study. AK collected the data. MT and AK wrote the code. MT wrote the original draft of the manuscript. Both authors read and approved the final manuscript.

Funding

The authors declare no source of funding.

Availability of data materials

The data described in this Data note can be freely and openly accessed on OSF at https://doi.org/10.17605/OSF.IO/AHCYU [9]. Please see Table 1 and reference list for details and links to the data.

Ethics approval and consent to participate

No human subjects were part of this study and permission was thus not required according to the Institutional Review Board guidelines of author one.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 31 October 2019 Accepted: 12 December 2019
Published online: 21 December 2019

References

1. Hicsonmez S, Sencar HT, Avcibas I. Audio codec identification from coded to transcoded audios. Digit Signal Process. 2013;23(5):1720–30.
2. Din M, Ratan R, Bhateja AK, Bhateja A. Multimedia classification using ANN approach. In Proceedings of the second International Conference on soft computing for problem solving (SocProS 2012), Dec 28–30, 2012, 2014 (pp. 905–910). Springer: New Delhi.
3. Asthana R, Verma N, Ratan R. Classification of distorted text and speech using projection pursuit features. In 2015 International Conference on Advances in computing, communications and informatics (ICACCI) 2015 Aug 10 (pp. 1408–1413). IEEE.
4. Maithani S, Din M. Speech systems classification based on frequency of binary word features. In 2004 International Conference on signal processing and communications, 2004. SPCOM'04. 2004 Dec 11 (pp. 193–197). IEEE.
5. Grajeda C, Breitinger F, Baggili I. Availability of datasets for digital forensics–And what is missing. Digit Invest. 2017;22:S94–105.
6. Fakouri R, Teimouri M. Dataset for file fragment classification of image file formats. BMC Res Notes. 2019;12:774. https://doi.org/10.1186/s13104-019-4812-0.
7. Mansouri Hanis F, Teimouri M. Dataset for file fragment classification of textual file formats. BMC Res Notes. 2019;12:801. https://doi.org/10.1186/s13104-019-4837-4.
8. VoxForg Speech Corpus [Internet]. http://www.voxforge.org/. Accessed 10 May 2019
9. Khodadadi A, Teimouri M. Audio File Fragments Dataset and Code [Internet]. OSF; 2019. https://doi.org/10.17605/OSF.IO/AHCYU.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.