INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

FAU Friedrich-Alexander-Universität

AUDIO LABS

Fraunhofer IIS

On the Effect of Coding Artifacts on Acoustic Scene Classification

Nagashree K. S. Rao and Nils Peters

International Audio Laboratories Erlangen, University of Erlangen-Nuremberg, Germany, {nagashree.s.rao, nils.peters}@fau.de

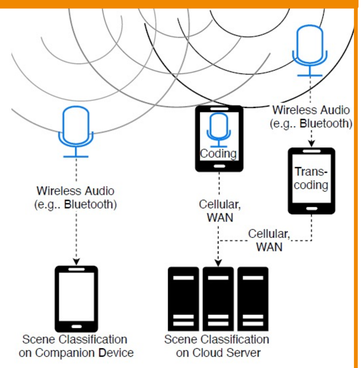
1. Introduction

State-of-the-art classifiers demand significant processing capabilities and memory for good performance which is challenging for resource-constrained mobile or IoT edge devices. It is more likely to deploy these models on more powerful hardware and classify audio recordings previously uploaded (or streamed) from low-power edge devices. In such a scenario, the edge device may apply perceptual audio coding to reduce the transmission data rate. This paper explores the effect of perceptual audio coding on classification performance using a DCASE 2020 challenge contribution [1]. We discuss how classification accuracy is degraded by lossy audio compression and a solution to improve it.

2. Problem Formulation

Perceptual audio coding:

- Significantly reduce the data rate of an audio signal .
- Does not intend to preserve the signal waveform.
- Aims to maintain a perceptually similar or even equal audio experience.



Existing state-of-the-art methods are trained & tested using uncompressed audio files.

Edge devices use audio coding, therefore its effects must be studied.

The authors of [1] have made source code and pre-trained models publicly available, which is used for our experiments.

We evaluate the pre-trained ensemble classifier using audio files perceptually encoded using ffmpeg . Results are shown in the table below

The classification accuracy dropped from 0.820 (classification accuracy on the uncompressed original evaluation files) to 0.352, a decrease by 57.1% .

Codec	Bit Rate [kbps]	Model Accuracy	Relative Decrease
None(Original)	1058	0.820	N/A
AAC	64	0.741	9.6%
MP3	64	0.724	11.7%
Opus	32	0.691	15.7%
HE-AAC	32	0.653	20.4%
SBC	64	0.632	22.9%
MP3	32	0.352	57.1%

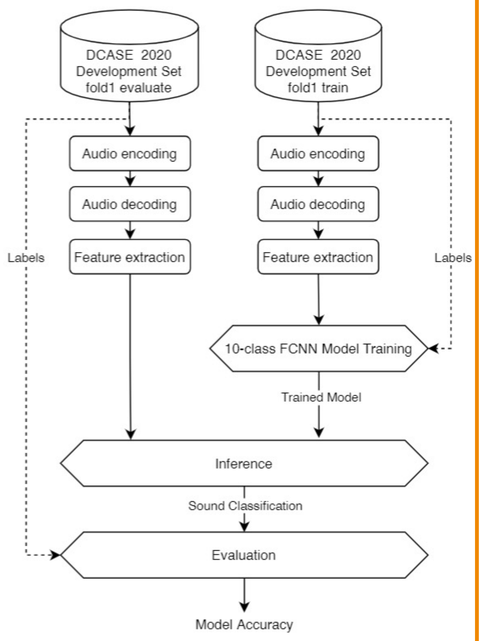
3. Our Solution

We propose retraining the model using additional data augmentation.

Augmentation data is generated by converting the training dataset into different encoded audio formats and bitrates.

We focus on the 10-class FCNN classification sub model.

We analysed the log-mel features across uncompressed and compressed audio data to find data that is most different than the training data.



Different audio codecs and bitrates that were chosen are: MP3 at 64kbps, AAC at 32kbps and HE-AAC at 16kbps and, 32kbps were selected.

Following are the training conditions:
Condition 1: The model is trained only with original audio data.
Condition 2: A Fully Augmented Dataset as mentioned in the original paper.
Condition 3: The model is trained with original audio data and augmentation data includes MP3 format coded audio data at 64kbps.
Condition 4 & **Condition 5:** Add HEAAC coded audio data at 16kbps and HEAAC coded audio data at 32kbps respectively, to training Condition 3.
Condition 6: Add AAC-coded audio data at 32kbps to training condition 4.
Condition 7: Add condition 6 to the Fully Augmented Data as in the original paper.

The newly trained models are evaluated with the same audio content as in problem formulation in the different conditions as follows:

- **No Coding Conditions:** Evaluation using the original evaluation data.
- **Seen Codec with seen bitrate Conditions:** The evaluation conditions use codecs at bit rates that were part of the training.
- **Unseen bitrate conditions:** The evaluation conditions feature the same codec used for training, but at different bit rates.
- **Unseen Codec Conditions:** The audio codecs that were not part of the model training are used for evaluation.

4. Performance Evaluation

Training Condition	Codec for Evaluation							
	None	AAC_{32}	$MP3_{64}$	$MP3_{32}$	AAC_{48}	AAC_{64}	$Opus_{64}$	SBC_{64}
1	0.703	0.458	0.550	0.261	0.477	0.525	0.587	0.391
2	**0.721**	0.558	0.615	0.301	0.573	0.622	0.638	0.555
3	0.668	0.566	0.635	0.249	0.602	0.630	0.556	0.363
4	0.696	0.631	0.662	0.447	0.638	0.651	0.587	0.534
5	0.699	0.630	0.666	0.319	0.648	0.663	0.593	0.508
6	0.697	**0.673**	0.675	0.561	0.664	0.671	0.610	0.560
7	0.720	0.670	**0.685**	**0.598**	**0.685**	**0.690**	**0.650**	**0.589**
relative performance increase from training condition 2 [%]	-0.1	20.1	11.4	98.7	19.6	10.9	1.9	6.1

No Coding Conditions: Re-training has no degrading effects on the classification accuracy. It improved gradually with inclusion of HE-AAC at 16kbps and AAC at 32kbps and further increased when the model was trained in the final training condition 7.
Seen Codec with bitrate Conditions (AAC at 32kbps, MP3 at 64kbps): The performance increased with the baseline training from 0.458 to 0.558 and from 0.550 to 0.615 respectively. The accuracy further improved with the addition of MP3 at 64kbps, and HE-AAC.
Unseen Bitrate conditions (MP3 at 32kbps, AAC at 48kbps, AAC at 64kbps)**:** The relative performance of classes MP3 at 32kbps, AAC at 48kbps, AAC at 64kbps improved by 98.7%, 19.6% and 10.9% respectively, as compared to the baseline performance.
Unseen codec Conditions (Opus at 64kbps, SBC at 64kbps)**:** The relative classification accuracy for Opus at 64kbps, SBC at 64kbps improved by 1.9% and 6.1% respectively, compared to the baseline performance.

5. Conclusions

- Perceptual compression artifacts can significantly degrade the accuracy of today's scene classification model.
- Data augmentation strategy for model training that includes perceptual audio coding improves robustness in classification even for audio codecs and/or bit rates not part of the model training.
- It does neither harm nor improve model performance when classification is performed on the original audio data.

References

[1] H. Hu et al., "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," DCASE2020 [Source Code:] https://github:com/MihawkHu/DCASE2020 task1
[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," Proceedings of the IEEE,
[3] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)