VPS VIBRATIONS IN PHYSICAL SYSTEMS

DOI: 10.21008/j.0860-6897.2022.3.02
Vibrations in Physical Systems, 2022, 33(3), 2022302

Lossy Coding Impact on Speech Recognition with Convolutional Neural Networks

Mateusz KUCHARSKI

Wrocław University of Science and Technology, Department of Computer Engineering,
Janiszewskiego 11/17, 50-372 Wrocław

Corresponding author: Mateusz KUCHARSKI, email: mateusz.kucharski@pwr.edu.pl

Abstract This paper presents research of lossy coding impact on speech recognition with convolutional neural networks. For this purpose, google speech commands dataset containing utterances of 30 words was encoded using four most common all-purpose codecs: mp3, aac, wma and ogg. A convolutional neural network was taught using part of the original files and later tested with the rest of the files, as well as their counterparts encoded with different codecs and bitrates. The same network model was also taught using mp3 encoded data showing the biggest loss in effectiveness of the previous network. Results show that lossy coding does have an effect on speech recognition, especially for low bitrates.

Keywords: lossy coding, convolutional neural networks, speech recognition.

1. Introduction

In recent years, speech and speaker recognition transitioned to the usage of neural networks, firstly to recurrent and recently convolutional methods [1,2,3,4]. While in commercial applications it is currently easier to use lossless or high-quality lossy coding, in forensics the recognition material tends to be limited, and in low bitrate lossy encoding [5]. In his previous works [6,7], author confirmed that lossy coding has a noticeable effect on formant parameters. Therefore, author decided to check, if it will have an impact on speech recognition.

The database used for this research was based on google speech commands dataset created by Pete Warden [8]. It consists of short (1 second) audio files containing 64 675 utterances of single words. They are 16 kHz, single channel wave files. In the used version, there are 30 common English words spoken by different speakers. The original dataset was converted into four most common all-purpose lossy codecs: mp3, aac, wma and ogg, with different bitrates, depending on the codecs' properties, as seen in Table 1. For easier and more consistent subsequent processing with a neural network, the created version of database was again converted back to wav files.

Table 1. Bitrate values for converted database.

	mp3	aac	wma	ogg
bitrate [kb/s]	16	32	32	16
	96	96	96	96
	160	192	192	-

For each set of converted as well as original wav data, 3000 files (100 per word) were separated for the purpose of evaluation of the trained network. This results in 12 sets (for wav and all conversions) containing 3000 files each. All of the separated files correspond to each other (separated files have the same IDs for wav and all conversions).

2. Research method

For the purpose of speech recognition, a simple convolutional neural network based on TensorFlow network [9] was used. For the training script, data is shuffled and split into training and validating sets. The network takes as input spectrogram images pre-processed from wav files [10]. Examples of spectrograms for all words in the database can be seen in Figure 1. The input is resized and normalized to fit the input of the neural network model. The model consists of two consecutive Conv2D layers, as well as a dense layer, all using rectified linear unit activation function [11]. The last layer outputs 30 possible labels with their probabilities for a given file. The evaluation script runs the same pre-processing for the separated files that were unused in the learning process, estimates their contents and checks those estimations with files' labels. The process starts with testing the original wav files and is repeated for all codecs and bitrates.

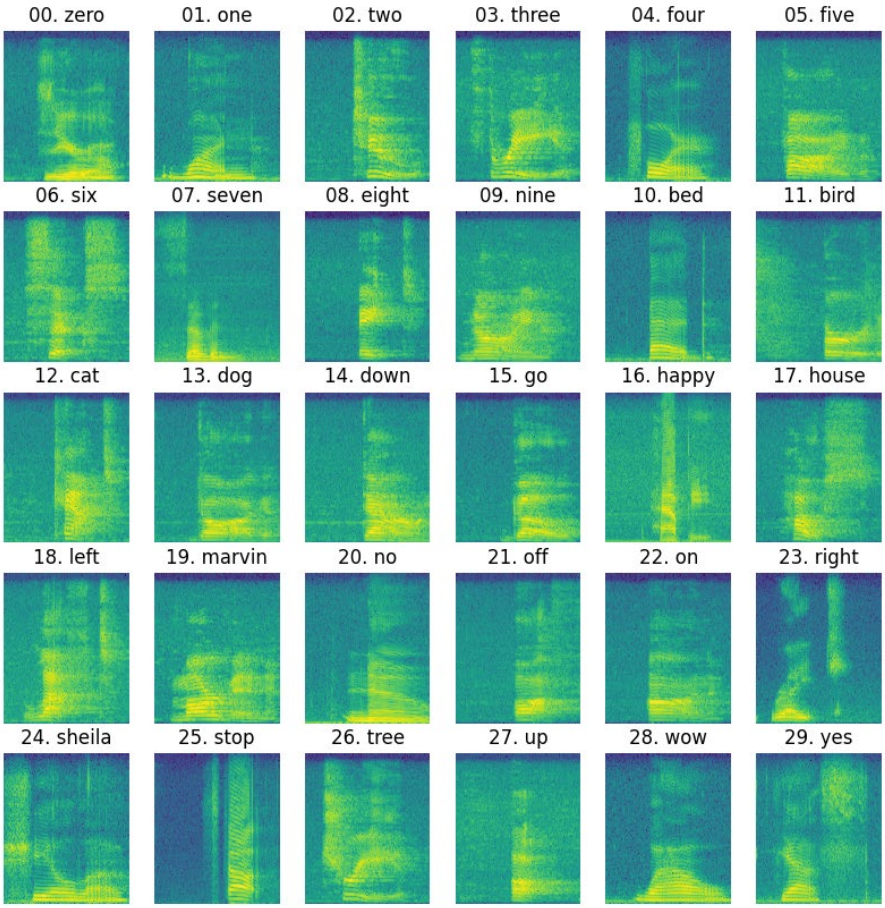

Figure 1. Examples of spectrogram images for all words in the database.

The results were evaluated with two metrics referred to later as "Confidence" and "Accuracy" [12]. For each tested file, the network returns a percentage likelihood of the file containing a word from each learnt label, all of which add up to 100%. The one with the highest percentage is taken as the answer. The mean of all predicted percentages for the label (no matter if it is actually the highest for all of the data) is presented as "Confidence" (it tends to be also described as accuracy). In author's evaluation "Accuracy" is the percentage of correctly classified files. It can be better understood with the following example: if for a file with label "happy" the network classifies it as being 49% "happy" and 51% "tree", it will have confidence of 49% and accuracy of 0% as the classification was wrong. For multiple files the accuracy is understood as a mean value. The author believes that using both of those metrics presents a better overview of the results than using just one of them. The results are presented in Table 2a. Confidence can also be presented using a confusion matrix. An example of a confusion matrix for the original wav data is shown in Figure 2 and for the lowest bitrate mp3 in Figure 3.

DOI: 10.21008/j.0860-6897.2022.3.02
Vibrations in Physical Systems, 2022, 33(3), 2022302

Table 2. Results for the model trained with: a) original wav data, b) lowest bitrate of mp3 coding.

a)

Coding	Bitrate [kb/s]	Confidence [%]	Accuracy [%]
wav	256	88	84
mp3	160	88	83
mp3	96	88	84
mp3	16	81	76
aac	192	86	81
aac	96	86	81
aac	32	86	81
wma	192	88	83
wma	96	88	83
wma	32	88	82
ogg	96	88	84
ogg	16	86	81

b)

Coding	Bitrate [kb/s]	Confidence [%]	Accuracy [%]
wav	256	81	76
mp3	160	81	76
mp3	96	81	76
mp3	16	86	81
aac	192	84	78
aac	96	84	78
aac	32	83	78
wma	192	82	78
wma	96	82	78
wma	32	82	75
ogg	96	81	76
ogg	16	81	76

3. Analysis of the results

As it can be seen, mp3, wma and ogg coding with high bitrates don't seem to affect the recognition. When looking at confidence all of them score exactly 88%, the same as the original wave files. The accuracy metric does waver slightly with a loss of 1 percentage point (which means about 30 more wrongly classified files than wav), however, it is still very good. The differences appear for low bitrates, as well as for all of the aac files, which all lose 2 percentage points of confidence and 3 points of accuracy. Mostly unaffected is the wma codec, however, it is worth noting that similarly to the aac its lowest bitrate is 32 kb/s, and not 16 kb/s like for ogg and mp3. The most apparent drop in network's effectiveness is for the low (16 kb/s) bitrate mp3 coding, with a loss of 7 points for confidence and 6 for accuracy, which means at least 180 wrongly classified files more than wav and other high bitrate codecs, including the mp3 itself.

Due to 16 kb/s mp3 files showing the biggest loss, the same network model was trained using this coding's equivalent of the original wav training dataset. The results are presented in Table 2b. The network trained this way shows a much clearer loss in recognition effectiveness. Testing data with the same coding as training data has the highest confidence and accuracy. However, they already are lower than for wav testing data with the previous network by 2 percentage points in confidence and 3 in accuracy. All other testing sets show worse results, they are also significantly worse than for the previous network.

Another aspect of speech recognition that should be pointed out is the existence of similar words. In the used dataset there are two such words: "three" and "tree". This can be best seen in Figures 2 and 3 containing confidence confusion matrixes for wav and mp3 testing sets, trained with original data. For wav the word "tree" was wrongly classified as "three" 20 times out of 100, making it the most wrongly classified word in the set. For the mp3 set the confidence classification is much worse, with 44 "trees" classified as "three", and with exactly 50 correctly classified "trees" when looking at the accuracy metric in Table 3b, what makes it overall very close to 50% of wrong classification between two subjects. It should be noted that this occurrence is not symmetric and for all sets, the word "three" has a classification rate similar to other words. This might signify that the network model prioritized learning this word over "tree", despite randomized input.

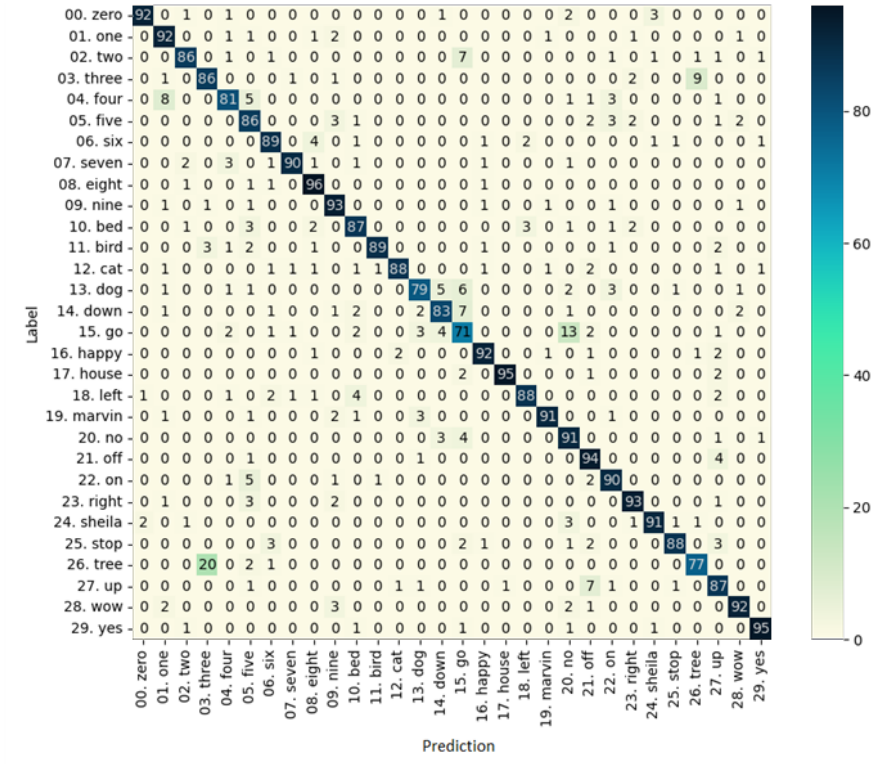

Figure 2. Confusion matrix of confidence for wav, trained on original data.

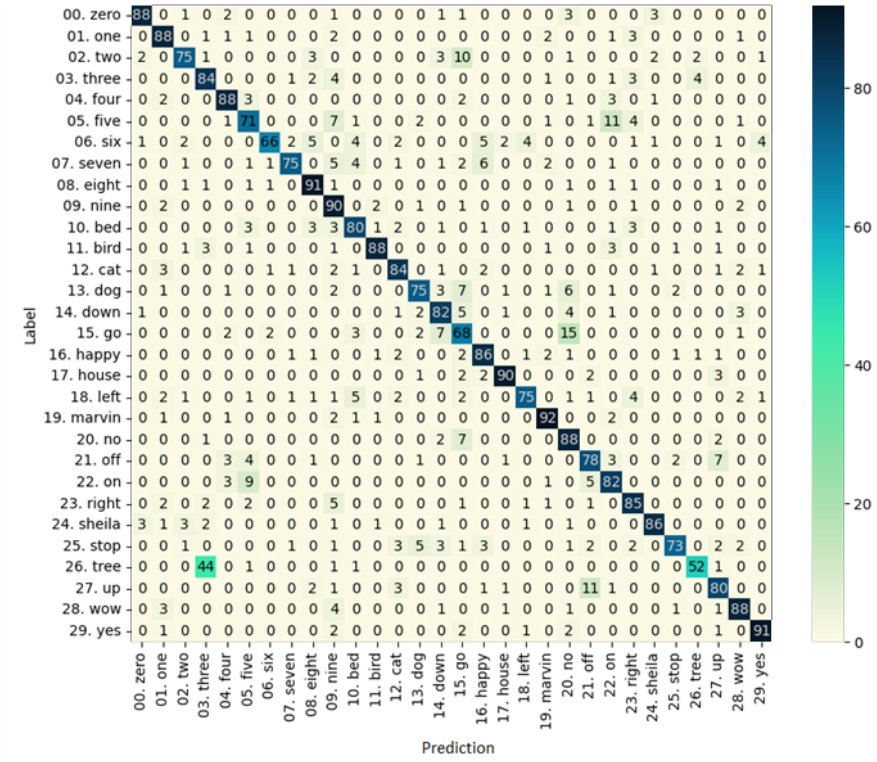

Figure 3. Confusion matrix of confidence for mp3, trained on original data.

DOI: 10.21008/j.0860-6897.2022.3.02
Vibrations in Physical Systems, 2022, 33(3), 2022302

Table 3. Example of accuracy distribution for: a) original wave files, b) lowest bitrate of mp3 coding.

a)

Word	Accuracy [%]	Word	Accuracy [%]
00. zero	88	15. go	64
01. one	90	16. happy	77
02. two	75	17. house	93
03. three	81	18. left	84
04. four	79	19. marvin	91
05. five	80	20. no	84
06. six	89	21. off	86
07. seven	85	22. on	81
08. eight	87	23. right	90
09. nine	92	24. sheila	87
10. bed	81	25. stop	83
11. bird	87	26. tree	76
12. cat	80	27. up	83
13. dog	74	28. wow	91
14. down	75	29. yes	95

b)

Word	Accuracy [%]	Word	Accuracy [%]
00. zero	85	15. go	57
01. one	85	16. happy	77
02. two	72	17. house	89
03. three	81	18. left	68
04. four	82	19. marvin	89
05. five	63	20. no	80
06. six	58	21. off	64
07. seven	67	22. on	76
08. eight	86	23. right	85
09. nine	88	24. sheila	83
10. bed	65	25. stop	68
11. bird	85	26. tree	50
12. cat	80	27. up	77
13. dog	73	28. wow	86
14. down	79	29. yes	81

4. Conclusions

Conducted experiments confirmed the hypothesis of low bitrate lossy coding having an impact on speech recognition effectiveness while using convolutional neural networks. While for medium and high bitrate values slight loss in effectiveness exists, it is not very significant. For the lower bitrate values, and especially for the mp3 codec with 16 kb/s bitrate, the results should be taken into consideration, as they displayed the loss of around 7 percentage points for overall recognition effectiveness. It was also shown that using this kind of data as training data for the neural network model can be detrimental, as model trained with this data was significantly worse in recognition of all of the data used for this research. Additionally words problematic due to their similarity, like “three” and “tree” proved much more problematic to distinguish when encoded, almost to the point of random classification.

It should be also noted that while the multimedia industry might not be too concerned about these results due to having the ability to use high bitrate and high-quality coding, forensic science often does not have this opportunity because of its nature and the evidence tending to be in low quality.

Because teaching the network using encoded files presented an increase in effectiveness for that particular coding, the next step in author’s research would be to test if implementing teaching neural networks with multiple codecs would increase resistance to efficiency loss due to lossy encoding.

Additional information

The author(s) declare: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

References

1. U. Kamath, J. Liu, J. Whitaker; Deep Learning for NLP and Speech Recognition; Springer Nature Switzerland AG 2019. DOI: 10.1007/978-3-030-14596-5
2. R.V. Pawar, P.P. Kajave, S.N. Mali; Speaker Identification using Neural Networks; Proceedings of World Academy of Science, Engineering and Technology Volume 7 August 2005
3. V. Delić, Z. Perić, M. Secujski, N. Jakovljević, J. Nikolić, D. Misković, N. Simić, S. Suzić, T. Delić; Speech Technology Progress Based on New Machine Learning Paradigm; Hindawi Computational Intelligence and Neuroscience Volume 2019. DOI: 10.1155/2019/4368036

4. O. Such, S. Barreda, a. Mojsej; A comparison of formant and CNN models for vowel frame recognition; 2019
5. H.A. Patil, A.E. Cohen, K.K. Parhi; Speaker Identification over Narrowband VoIP Networks; Forensic Speaker Recognition, Springer: New York, 2012. DOI: 10.1007/978-1-4614-0263-3_6
6. M. Kucharski, S. Brachmański; Coding Effects on Changes in Formant Frequencies in Japanese Speech Signals, Vibrations in Physical Systems 2019, 1, 30, 243-250
7. M. Kucharski, S. Brachmański; Coding effects on changes in formant frequencies in Japanese and English speech signals; EURASIP Journal on Audio, Speech and Music Processing, 2022, submitted
8. P. Warden; Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition; arXiv:1804.03209, 2018. DOI: 10.48550/arXiv.1804.03209
9. Simple audio recognition: Recognizing keywords; https://github.com/tensorflow/docs/blob/master/site/en/tutorials/audio/simple_audio.ipynb (access 28.04.2022)
10. A.B. Downey; Think DSP: Digital Signal Processing in Python; Version 1.1.4, Green Tea Press, 2014.
11. TensorFlow Core v2.9.1 API documentation for Python: https://www.tensorflow.org/api_docs/python/tf (access 15.08.2022)
12. C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger; On Calibration of Modern Neural Networks; International Conference on Machine Learning, 2017. DOI: 10.48550/arXiv.1706.04599