Audio Signal Mapping into Spectrogram-Based Images for Deep Learning Applications

Dejan Ćirić, Zoran Perić, Jelena Nikolić, Nikola Vučić
Department of Telecommunications
Faculty of Electronic Engineering, University of Niš,
Niš, Serbia
dejan.ciric@elfak.ni.ac.rs; zoran.peric@elfak.ni.ac.rs;
jelena.nikolic@elfak.ni.ac.rs; nikola.vucic@elfak.ni.ac.rs

2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH) | 978-1-7281-8229-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/INFOTEH51037.2021.9400698

Abstract—**Various features generated from raw audio signals can be used as an input of a deep learning model. They include hand-crafted features such as mel-frequency cepstral coefficients, two-dimensional time-frequency representations and raw audio data. In most cases, the time-frequency representations are related to so-called spectrogram-based images. Having an image at the deep learning input enables to apply performance improvement accumulated in video and image processing. However, spectrogram-based images have some specific properties that should be taken into account when a deep learning model is designed. This paper deals with mapping of audio signals into the most common spectrogram-based images. Some unique properties of these images as well as the way how they are generated are analyzed here for a particular case of fridge sounds.**

Keywords-deep learning; audio signal; time-frequency image; spectrogram; gammatonegram; constant Q transform

I. Introduction

Performance of deep neural networks (DNNs) has reached a rather high level outperforming many other audio signal processing methods such as Gaussian mixture models, hidden Markov models and non-negative matrix factorization [1]. A question that arises when deep learning model is to be designed is how to represent the data. While the answer is rather obvious in images, this is not the case in audio. Here, a variety of representations have been used including hand-crafted features, machine discovered features, spectrogram-based images and raw audio data [2]. It is worth mentioning that there is no consensus on what input representation to DNN should be used [1]. In spite of the fact that hand-crafted features are well designed, due to knowledge bias and high compression ratio all useful information can hardly be retained.

In order to avoid usage of hand-crafted features and relying on a designed filter bank, so called data-driven statistical model learning has recently been proposed, where data-driven filters are learned and used [1]. With this approach, raw waveform representation of audio signal is applied as a lossless input, see, for example event classification in SoundNet or WaveNet [2]. While usage of raw audio data avoids hand-crafted features and allows exploitation of the improved modeling capability of deep learning, it also incurs higher computational costs and data requirements [1]. An alternative approach is to use either a full or reduced resolution spectrogram-based images as an input to deep learning model. As a more compact representation of an input, these images require less data and training to achieve the same classification performance as a method using raw audio data [1].

Spectrogram-based images can be considered as a compromise between hand-crafted features and raw data, and usage of such features is also more in line with human perception. They typically retain more information than hand-crafted features, and have lower dimensionality than raw data [2]. However, it has also been highlighted that they can be too generic and fail to describe specific content of sound [3]. In acoustic event detection, it has been shown that high resolution spectrograms have a significant advantage over hand-crafted features [4]. Such spectrograms incorporate complexity not only in the frequency domain, but also in the form of wide range of temporal structures [4]. Since spectrogram is a sort of image, that is, a very detail and accurate visual representation of audio [5], it can be processed as an image performing neural style transfer with DNN. In this way, the performance gains from vision models are translated to the audio domain [2]. At the same time, spectrogram has some unique features due to which the results of DNN processing are not at the same level as for visual images [1, 2, 6].

Different two-dimensional time-frequency representations of audio signals have been used as an input for DNN classifiers. They include the following images: spectrogram, mel-spectrogram, cochleagram, gammatonegram, constant Q transform (CQT) image, chromagram, tempogram, auditory image map (AIM), stabilized auditory image (SAI), etc. Many of these representations are also called spectrogram-based images in the literature, since the most common representation is spectrogram, and these images are visually similar to a spectrogram. Some of these images have more often been used, such as spectrogram and mel-spectrogram, while some other have rarely been applied in practice, such as AIM and SAI [7].

This paper reviews the most important spectrogram-based images and issues arising from their usage for generating neural networks in audio applications. Lin-power and log-power spectrograms, mel-spectrograms, gammatonegrams

This research was supported by the Science Fund of the Republic of Serbia, 6527104, AI-Com-in-AI.

978-1-7281-8229-2/21/$31.00 ©2021 IEEE

(cochleagrams), CQT images and chromagrams are generated from sounds of two fridges of different qualities, samples A and B. In this way, the visual representations of sounds of these fridges can be compared. The effects of using both linear and logarithmic frequency scale and different color maps are observed. The paper is organized as follows: Section II briefly describes the spectrogram-based images. Method of investigation is explained in Section III, while the results are presented in Section IV. Conclusion and future work are summarized in Section V.

II. Spectrogram-Based Images

Spectrogram is a two-dimensional representation of audio frequency content over time, that is, spectrogram is temporal sequence of spectra. Brightness or color represents the strength of a frequency component at each time frame [2]. DNN can learn from specific spectro-temporal shapes and make classification with significant level of robustness [4].

Other two-dimensional time-frequency representations (images) share a common property of presenting frequency content over time with spectrogram. However, the frequency content and the method of calculation are specific for each particular image. In comparison to classical spectrogram, CQT and mel-spectrograms reduce the output quality. An interesting situation often met in practice is the one with multiple simultaneous (overlapped in time) sound events. In such a case, the most prominent sound properties in some spectrogram-based images, such as mel-spectrogram are diluted. On the other hand, this is not the case in high resolution spectrograms where these properties are still identifiable [8].

A. Spectrogram

Audio signal is transformed into spectrogram with the short-term Fourier transform (STFT) yielding complex values as a result. The *f*-th component of discrete Fourier transform (DFT) of the *t*-th frame of signal *x*[*n*] is calculated as

$$X[t,f]=\sum_{k=0}^{N-1} w[k]x[tN+k]e^{\frac{-i2\pi kf}{N}} , \qquad (1)$$

where *w*[*k*] represents a window function, e.g., Hamming or Blackman, which is used to enforce continuity and periodicity at the edge of frames. In (1), the shift or hop size between consecutive frames is equal to the frame size (*N*) meaning that there is no overlap between frames. However, the hop size used in practice is often smaller than the frame size allowing smoother STFT and introducing statistical dependencies between frames [3]. A spectrogram is generated from the STFT (complex) results as a matrix, where each column is the DFT result of a particular signal frame. Typically, magnitude of these complex results is used for plotting the spectrogram.

B. Mel-spectrogram

Instead of having all frequencies from the DFT separately, it can be useful to analyze the content, e.g., energy or energy ratios present in particular frequency bands. These bands can be equally spaced on the frequency axis or unequally according to a certain law, such as logarithmic law or perceptual law [3]. In that regard, the shape of the applied filter, the number of bands and their overlap can be different. An example is usage of a filter bank with frequencies aligned to the Bark scale [9].

Human ear is not equally sensitive to different frequencies, and its resolution varies along the frequency axis. In order to analyze sound in a similar way as done in human auditory system, it is better to use non-linear frequency scales. Mel scale is the frequency scale matching perceptual behavior, that is, it is related to the psychological sensation of heights of a pure sound [3]. A relation between the mel scale, *mel*(*f*), and frequency scale in Hertz, *f*, can be given as

$$mel(f)=\frac{1000}{\log(2)}\log\left(1+\frac{f}{1000}\right). \qquad (2)$$

Here, two pairs of mel-frequencies with a distance of delta are perceived as equidistant ones. Mel-spectrogram is obtained by applying a bank of overlapping triangular filters calculating the energy of spectrum in each band (see Fig. 1). The filter bandwidth increases with frequency.

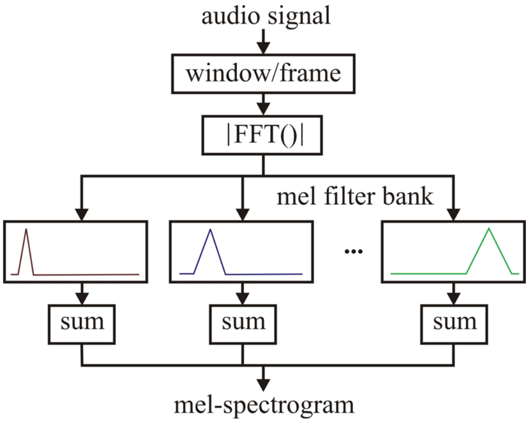

Figure 1. Obtaining mel-spectrogram from audio signal, where fast Fourier transform is denoted by FFT.

C. Gammatonegram (Cochleagram)

Gammatonegram, which is also known as cochleagram, is a relatively new time-frequency representation of an audio signal in the machine hearing field. It is derived from gammatone warping function, that is, the gammatone auditory filter function [7]. The idea with this function is based on fitting experimental observations to mammalian cochlea frequency selectivity providing in this way bio-mimetic auditory features. This function was originally applied to one-dimensional gammatone cepstral coefficients and such an approach outperformed other one-dimensional features [10].

Gammatone filters are linear filters having specific (gammatone) impulse response given by

$$g(n)=an^{P-1}\cos(2\pi f_c n+\phi)e^{-2\pi bn} , \qquad (3)$$

where a is amplitude, n is discrete time, P is the filter order, ϕ is the phase shift of the carrier (related to the position of the envelope on the carrier), f_c is the frequency of the carrier (related to the central frequency of the filter, in kHz), and b is the temporal decay coefficient (related to the filter bandwidth) [3, 7]. The gammatone impulse response is composed of a sinusoidal carrier wave (tone) modulated in amplitude by an envelope having the same form as a scaled gamma distribution function [3]. The bandwidth of gammatone filter can be obtained as $b = 1.019 \times f_{erb}$, where f_{erb} represents the equivalent rectangular bandwidth (ERB) scale, given as $f_{erb} = 24.7 \times (4.37 \times f_c + 1)$ [7]. Gammatonegram can be obtained by stacking the gammatone filter output vectors from each frame of an audio signal.

Here, the gammatone auditory filters are implemented in the same way as in [11]. The parameters of the gammatonegram calculation are: number of channels of gammatone auditory model filter-bank, the lowest and the highest frequency as well as the frame and hop size. The output energy of each band is integrated over frames of the defined size with the defined hop size.

D. Constant Q Transform

Performance of the CQT is well-known in music analysis [12], e.g., for finding correspondence between filters and musical notes by identifying appropriate center frequencies. Thus, CQT in combination with a refined frequency estimation observing phase changes is used for tone center analysis [13]. Its perceptual relevance has already been presented.

CQT can be described as a bank of filters, which is similar to the Fourier transform. However, the filter center frequencies are here geometrically spaced $f_k = f_0\, 2^{k/b}$, while the ratio of frequency to filter bandwidth, „Q", is constant [13]

$$Q = \frac{f_k}{\Delta_k^{cq}} = \left(2^{\frac{1}{b}} - 1\right)^{-1} \quad (k = 0, \ldots) \ . \tag{4}$$

The number of filters per octave is determined by the variable b. The appropriate window length should individually be chosen for each frequency bin of the CQT. In order to have an integer value of Q, k-th CQT bin should be Q-th bin of the DFT with window length $Q\,(f_s / f_k)$. Calculation steps include the following [13]: minimal frequency f_0 and the number of bins per octave b should be chosen. The values of K and N_k are calculated as

$$K := \left\lceil b \log_2\left(\frac{f_{\max}}{f_0}\right)\right\rceil, \quad Q = \left(2^{\frac{1}{b}} - 1\right)^{-1}, \tag{5}$$

$$N_k := \left\lceil Q \frac{f_s}{f_k} \right\rceil, \quad (\text{for } k < K) \, . \tag{6}$$

The k-th CQT bin is obtained as

$$N_k^{-1} \sum_{n<N_k} x[n]\, \omega_{N_k}[n]\, e^{-2\pi i n \frac{Q}{N_k}} \ . \tag{7}$$

It is worth noting that magnitudes of CQT over an array of overlapping windows form a pyramid, as a consequence of frequency dependent resolution, that is, resolution varies with frequency [7]. This means that frequency bin sizes of the CQT span from a single pixel to a range of pixels. In order to make comparison and feature processing easier, the mentioned pyramid is rasterised resulting in a rectangular CQT feature. Here, its dimensionality is the same as the one of spectrogram.

III. Method of Investigation

The spectrogram-based images lin and log-power spectrogram, mel-spectrogram, gammatonegram, CQT spectrogram and chromagram are calculated according to the previously described procedures implemented in the Python code. A fridge is chosen as a representative of industrial product that we use in this phase of the research. In order to have two distinctive categories (classes), sounds of two fridges having clear difference between each other are used for calculation of the time-frequency images. Audio signals containing these sounds have length of 5 s, and the sampling frequency is 48 kHz. They are shown in Fig. 2, while their spectra are given in Fig. 3.

As one can notice from Fig. 2, in the time domain, the most prominent difference between the given sounds is in the low frequency modulation present in the sample B. Here, sample B has worse quality of the product that is also reflected in worse sound from the perceptual standpoint. Regarding the sound spectra, general shape of spectra is similar for both samples, but there is an obvious shift in the positions of prominent peaks (harmonics). In other words, sample B has a shift of corresponding peaks towards higher frequencies.

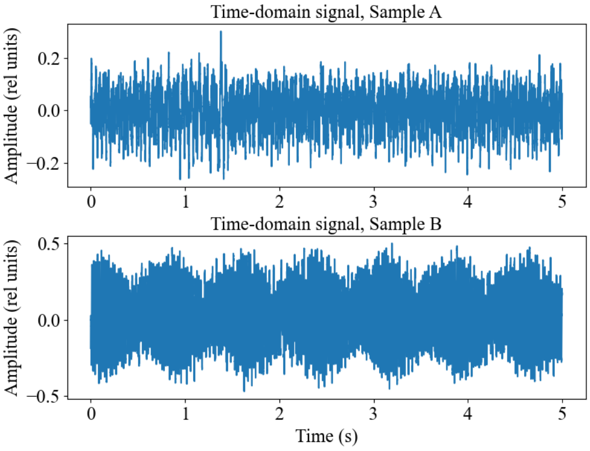

Figure 2. Audio signals of two fridges (samples A and B) in time domain.

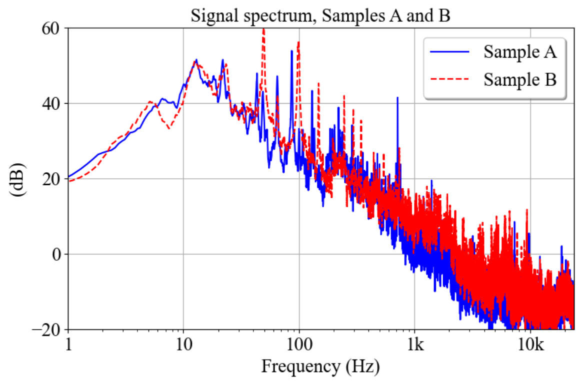

Figure 3. Spectra of sounds of two fridges (samples A and B).

IV. Obtained Spectrogram-Based Images

Classical spectrogram is calculated by applying the STFT having the frame size of 2048 samples (approximately 43 ms at the sampling rate of 48 kHz) and hop size of one quarter of the frame size, that is, 512 samples or about 11 ms. The lin-power spectrogram is obtained taking the magnitude of the obtained complex results of the STFT (see Fig. 4). The mentioned low frequency modulation present in the fridge sample B is clearly visible in the given spectrogram at frequencies below approximately 100 Hz. Other differences between the sounds of fridge samples A and B are hardly visible.

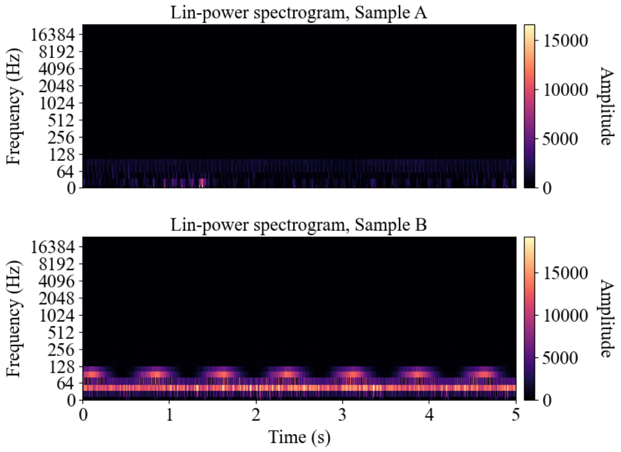

Figure 4. Lin-power spectrograms of two fridges (samples A and B) obtained by STFT using the frame size of 2048 samples and the hop size of 512 samples (frequency axis is logarithmic).

When absolute power of the lin-power spectrogram is transformed into dBs, log-power spectrogram is generated (see Fig. 5). The dynamic range of values of spectrogram is now reduced in comparison to lin-power spectrogram and some other differences between spectrograms for fridge samples A and B become more prominent, such as differences in amplitude values of some harmonics. The low frequency modulation of the fridge sample B is still clearly noticeable.

The mel-spectrograms for analyzed audio samples obtained using the same frame and hop size, 2048 and 512 samples, respectively, and 96 mel-bands are shown in Fig. 6. It is worth mentioning that the frequency resolution is affected by the number of used mel-bands. Here, in spite of using rather large value of mel-bands, the frequency resolution is reduced in comparison to spectrograms. Consequently, some details of the differencies in the frequency domain between the images for fridge samples A and B are blured. Moreover, the low frequency modulation in the fridge sample B becomes less prominent.

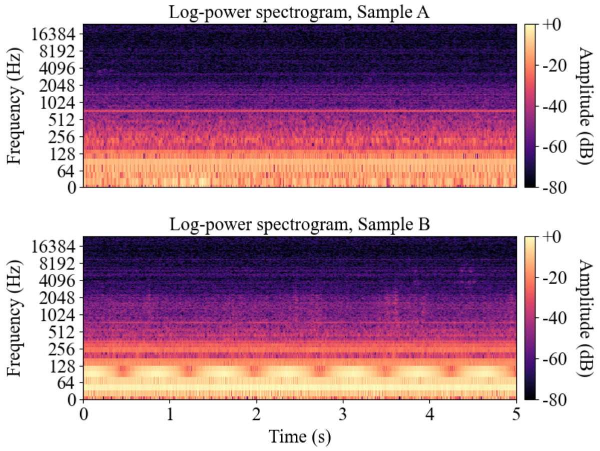

Figure 5. Log-power spectrograms of two fridges (samples A and B) obtained using the frame size of 2048 samples and the hop size of 512 samples (frequency axis is logarithmic).

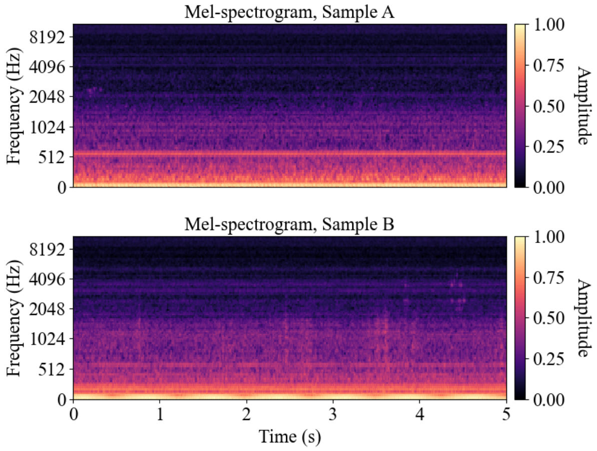

Figure 6. Mel- spectrograms of two fridges (samples A and B) obtained using the frame size of 2048 samples and hop size of 512 samples for the number of mel-bands of 96.

In order to generate the gammatonegram, the audio signal is filtered by 64 channel gammatone auditory model filter-bank, where the lowest frequency is 20 Hz and the highest frequency is half of the sampling frequency. As for other spectrogram-based images, the frame and the hop size are also 2048 samples and 512 samples, respectively. Patterns of gammatonegrams are similar to those of the spectrograms, although the resolution in frequency is here reduced, and there are certain differences in the values of frequency bins as a consequence of different methods of calculation (see Fig. 7). Again, we can notice that low frequency modulation in the fridge sample B is prominent.

CQT power spectrograms of the analyzed two samples of fridges are presented in Fig. 8. Comparing these images to spectrograms or gammatonegrams, the pattern is now somewhat different. There are visible horizontal lines indicating more prominent frequency content at these frequency bins. In addition, prominent differences between the patterns of CQT images for fridge samples A and B can be noticed. It is worth mentioning that these differences are not the same as those present in the spectrograms, mel-spectrograms and gammatonegrams. This could be an indication that CQT spectrogram is able to provide additional information not clearly distinguishable in the other three spectrogram-based images.

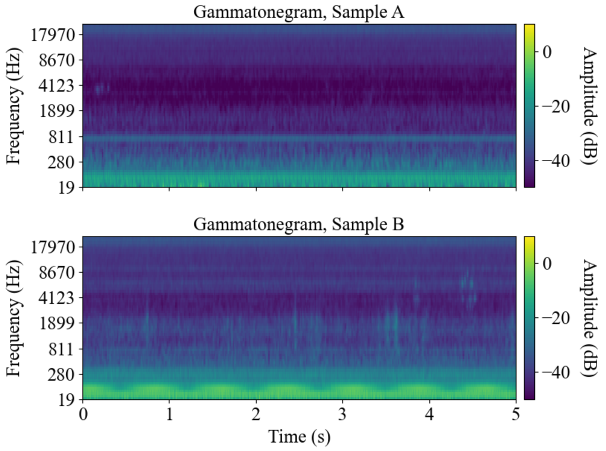

Figure 7. Gammatonegrams (cochleagrams) of two fridges (samples A and B) obtained using the frame size of 2048 samples, hop size of 512 samples and 64 channels of gammatone auditory model filter-bank.

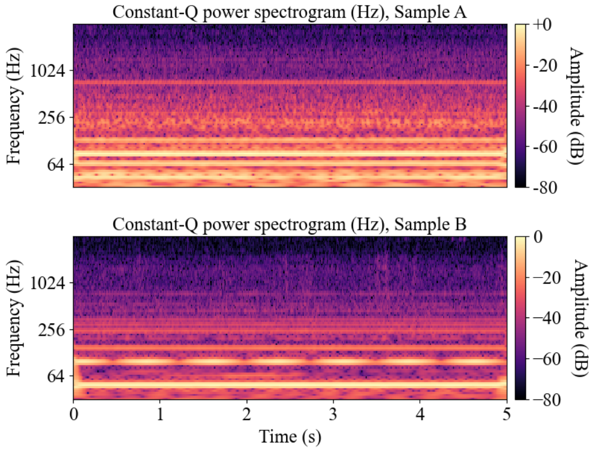

Figure 8. CQT power spectrrograms of two fridges (samples A and B) obtained using the hop size of 512 samples.

The chromagrams are here calculated using the STFT and applying the same values of the frame size and hop size as well as 12 chroma bins (see Fig. 9). The frequency resolution is rather coarse. However, some prominent features of the analyzed sounds can be seen. What is even more important, the chromagram pattern for the fridge sample A is rather different from the pattern for the sample B. It is interesting to note that low frequency modulation of the fridge sample B generates specific low frequency shape of its chromagram.

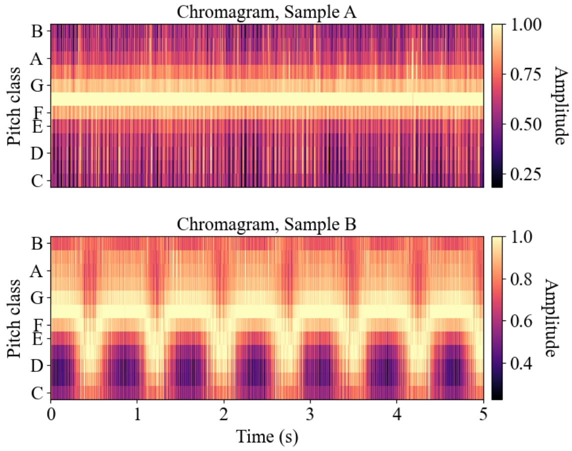

Figure 9. STFT chromagrams of two fridges (samples A and B) obtained using the frame size of 2048 samples, the hop size of 512 samples and 12 chroma bins.

An illustration of changing the image presentation by changing the logarithmic frequency axis with the linear one is given through the log-power spectrograms in Fig. 10. Most of the important frequency components of the analyzed fridge sounds, such as harmonics and low frequency modulation, appear in the frequency range up to approximately 1 kHz. Importance of these components can be recognized, but it is not possible to make a distinction between them and to see some particular properties such as low frequency modulation in the log-power spectrogram shown on linear frequency axis.

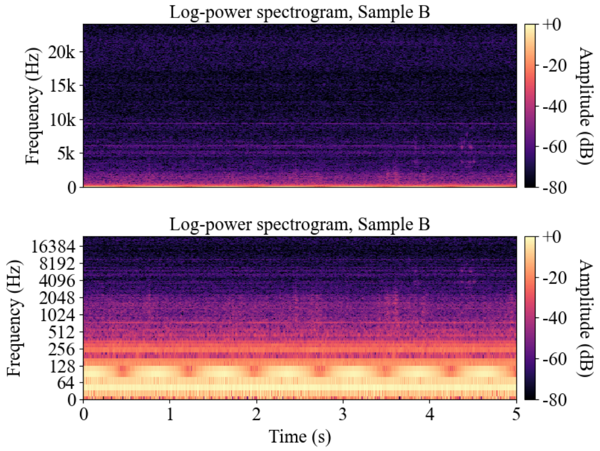

Figure 10. Log-power spectrograms of fridge sample B obtained using the frame size of 2048 samples and the hop size of 512 samples: frequency axis is linear (top) and logarithmic (bottom).

Regarding visualization of spectrogram-based images, they can be plotted using different parameters including different color maps. A number of maps are available in Python plotting functions. They are typically split into categories, such as sequential, diverging, cyclic or qualitative [14]. The choice of a color map affects how the presented data are perceived. For example, perceptually uniform color map can be a good choice since in this color map equal steps in data are perceived as equal steps in the color space [14]. One of those maps belonging to category sequential is “magma” used for previously shown figures including Fig. 5. Log-power spectrograms of the fridge sample B generated using three other color maps (“gist_ncar” - sequential 2, “coolwarm” - diverging and “blues” - sequential) are given in Fig. 11.

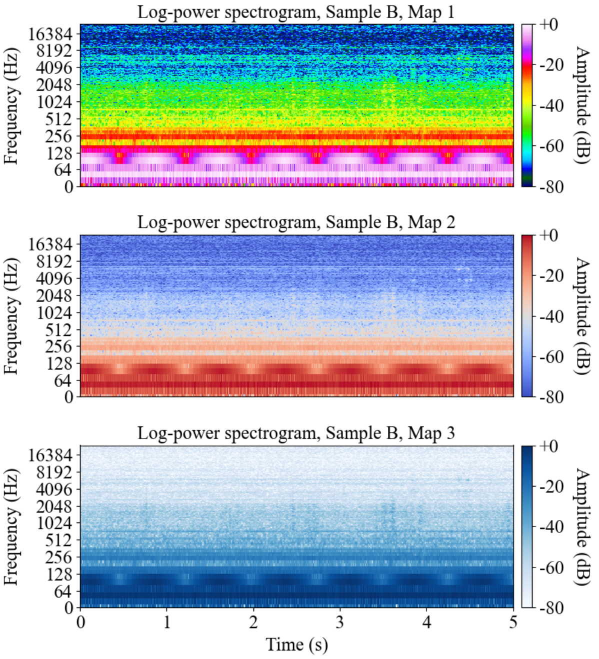

Figure 11. Log-power spectrograms of fridge sample B obtained using the frame size of 2048 samples and the hop size of 512 samples as well as different color maps: “gist_ncar” (top), “coolwarm” (middle) and “blues” (bottom).

V. Summary and Conclusion

Mapping of audio signals into two-dimensional time-frequency representations for the purpose of using them at the deep learning input has been exploited more and more recently. These representations are often called spectrogram-based images as they are similar to a classical spectrogram. Various such images are applied in deep learning applications, where the most common ones are spectrogram, mel-spectrogram, gammatonegram (cochleagram), CQT spectrogram and chromagram. Calculation of these images and their basic properties are analyzed in this paper. Audio signals of two fridges belonging to different categories according to their quality and generated sound are used as an input for mapping.

We have noticed that the obtained spectrogram-based images have some unique properties, but also that a common pattern can be found in many of them. Specifically, the most obvious properties of the time-frequency representations in the analyzed cases of two fridges are related to appearance of specific harmonics and low frequency modulation present in the fridge of worse quality. The visibility of these properties depends on the particular algorithm applied for calculation of certain spectrogram-based image, but also on the frequency resolution of the image. Thus, in some of the images such as mel-spectrogram, the frequency resolution is decreased in comparison to the case of full-resolution spectrogram. Based on the presented results, it can be concluded that some of the spectrogram-based images such as spectrogram, mel-spectrogram and gammatonegram have a rather similar pattern, while CQT image shows the information in a bit different way.

In order to obtain quantitative conclusions regarding the information contained in the spectrogram-based images, a quantitative analysis of these images will be carried out in the very next phase of this research. It will include sounds of more industrial products such as DC motors and compressors. The analysis will be extended to deal with other image parameters such as time-frequency resolution and down-scaling of a full resolution image into lower resolution image. Dependence of performance of particular deep learning model on the input spectrogram-based image will be studied as well. Eventually, effects of applying the advanced methods of quantization to the deep learning model input on the model performance will be investigated.

References

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, “Deep learning for audio signal processing,” J. Selected Topics Sig. Proc., vol. 13, no. 2, pp. 206–219, May 2019.

[2] P. Verma, J. O. Smith, “Neural style transfer for audio spectrograms,” 31st Conf. Neural Inf. Proc. Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[3] T. Virtanen, M. D. Plumbley, D. Ellis, Computational Analysis of Sound Scenes and Events, Springer: New York, 2018, Chapter 4: R. Serizel, V. Bisot, S. Essid, and G. Richard, Acoustic Features for Environmental Sound Analysis pp. 71–101.

[4] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, “Exploiting spectro-temporal locality in deep learning based acoustic event detection,” EURASIP J. Aud., Speech, Music Proc., article no. 26, 2015.

[5] Y. Zeng, H. Mao, D. Peng, and Z. Yi, “Spectrogram based multi-task audio classification,” Multimedia Tools App., vol. 78, no. 3, pp. 3705–3722, 2019.

[6] D. Rothmann, “What’s wrong with CNNs and spectrograms for audio processing,” https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd.

[7] I. McLoughlin, Z. Xie, Y. Song, H. Phan, and R. Palaniappan, “Time–frequency feature fusion for noise robust audio event classification,” Circuits, Systems, Sig. Proc., vol. 39, pp. 1672–1687, 2020.

[8] M. Espi, M. Fujimoto, and T. Nakatani, “Acoustic event detection in speech overlapping scenarios based on high-resolution spectral input and deep learning,” IEICE Trans. Inf. Syst, vol. E98-D, no. 10., Oct. 2015.

[9] S. Böck, A. Arzt, F. Krebs, and M. Schedl, “Online real-time onset detection with recurrent neural networks,” 15th Int. Conf. Dig. Aud.Effects (DAFx-12), York, UK , 17-21 Sep., 2012.

[10] J.W. Dennis, Sound Event Recognition in Unstructured Environments Using Spectrogram Image Processing. Ph.D. thesis, Nanyang Technological University, Singapore, 2014.

[11] D. P. W. Ellis, “Gammatone-like spectrograms,” http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/.

[12] A. Lacoste, and D. Eck, “A supervised classification algorithm for note onset detection,” EURASIP J. Adv. Sig. Proc., vol. 2007, article ID. 43745, 13 pages, 2007.

[13] H. Purwins, B. Blankertz and K. Obermayer, “A new method for tracking modulations in tonal music in audio data format,” *IEEE-INNS-ENNS Int. Joint Conf. Neural Net. IJCNN 2000. Neural Comp.: New Challenges Persp. New Mill.*, Como, Italy, vol. 6, pp. 270-275, 2000.

[14] “Choosing colormaps in Matplotlib,” https://matplotlib.org/tutorials/colors/colormaps.html.