Daniel Rothmann Follow
Developer @ Kanda. Especially interested in AI, VR and all things transformative.
Mar 25 · 8 min read

(Photo credit: Jack Hamilton)

What's wrong with CNNs and spectrograms for audio processing?

In recent years, great results have been achieved in generating and processing images with neural networks. This can partly be attributed to the great performance of deep CNNs to capture and transform high-level information in images. A notable example of this is the process of image style transfer using CNNs proposed by L. Gatys et. al. which can render semantic content of an image in a different style [1].

The process of neural style transfer is well explained by Y. Li et. al: *"this method used Gram matrices of the neural activations from different layers of a CNN to represent the artistic style of a image. Then it used an iterative optimization method to generate a new image from white noise by matching the neural activations with the content image and the Gram matrices with the style image"* [2].

In simpler terms, these results can be thought of as being achieved by generating images according to combinations of features from source content and style images at different levels of abstraction. As an

example, this could be maintaining the high level structures and contours of the content image while incorporating colors and lower level texture features of the style image.

An example of the transfer of style features ("B") onto a content image ("A") by L. Gatys et. al.

The performance of style transfer in the realm of visual processing has been quite impressive and lends itself to optimism for "smarter" audio processing algorithms if similar results can be achieved. Since spectrograms are two-dimensional representations of audio frequency spectra over time, attempts have been made in analyzing and processing them with CNNs. It has been shown, that it is possible to process spectrograms as images and perform neural style transfer with CNNs [3] but, so far, the results have not been nearly as compelling as they have been for visual images [4].

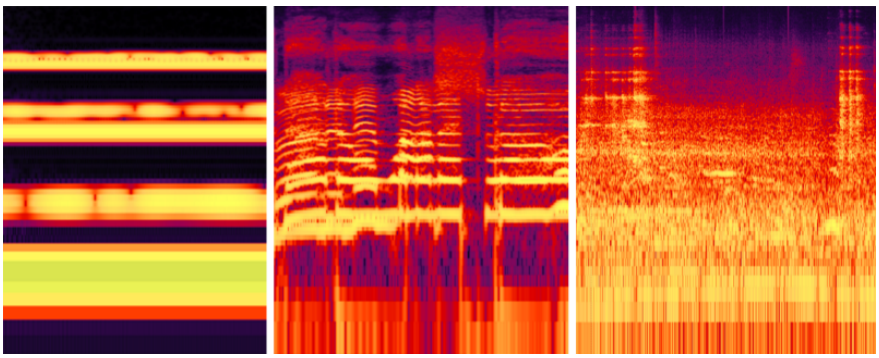
To overcome this challenge and to produce better results in neural audio processing we might then need to consider why style transfer with CNNs do not perform as well on spectrograms. **Essentially, these techniques apply machine vision in order to do machine hearing.** I believe this poses an essential problem which could be hindering the progress of AI-assisted technologies in audio processing. While the problem can undoubtedly be approached from many angles, it might be worthwhile exploring the differences between images and spectrograms and, as a consequence, some of the differences between seeing and hearing.

Sounds are "transparent"

One challenge posed in the comparison between visual images and spectrograms is the fact that visual objects and sound events do not accumulate in the same manner. To use a visual analogy, one could say

that sounds are always "transparent" [4] whereas most visual objects are opaque.

When encountering a pixel of a certain color in an image, it can most often be assumed to belong to a single object. Discrete sound events do not separate into layers on a spectrogram: Instead, they all sum together into a distinct whole. That means that a particular observed frequency in a spectrogram cannot be assumed to belong to a single sound as the magnitude of that frequency could have been produced by any number of accumulated sounds or even by the complex interactions between sound waves such as phase cancellation. This makes it difficult to separate simultaneous sounds in spectrogram representations.



Three examples of difficult scenarios of spectrogram analysis. (Left): Two similar tones cause uneven phase cancellations across frequencies. (Middle): Two simultaneous voices with similar pitch are difficult to tell apart. (Right): Noisy and complex auditory scenes make it particularly difficult to distinguish sound events.

The axes of spectrograms do not carry the same meaning

CNNs for images use two-dimensional filters that share weights across the x and y dimensions [4]. As earlier described, this builds on the assumption that features of an image carry the same meaning regardless of their location. For this to be true, you should also assume that the x and y axes of the data have the same implications to the meaning of the content. For example, a face is still a face regardless of whether it is moved horizontally or vertically in an image.

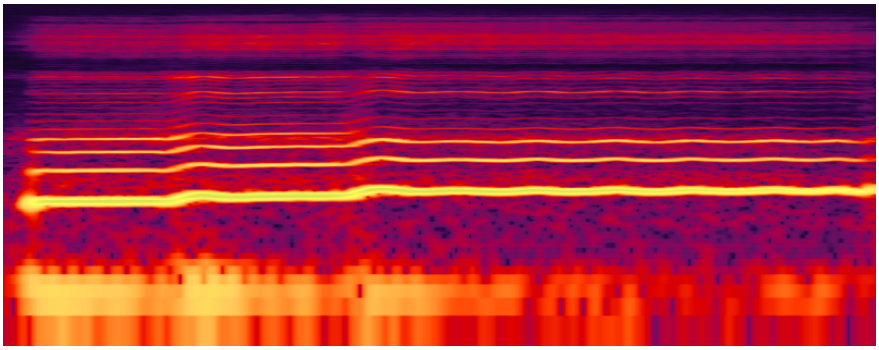
In spectrograms, the two dimensions represent fundamentally different units, one being strength of frequency and the other being time. Moving a sound event horizontally offsets its position in time and it can

be argued that a sound event means the same thing regardless of when it happens. However, moving a sound vertically might influence its meaning: **Moving the frequencies of a male voice upwards could change its meaning from man to child or goblin**, for example. Performing a frequency shifts of a sound event can also changes its spatial extent [4]. Therefore, the spatial invariance that 2D CNNs provide might not perform as well for this form of data.

The spectral properties of sounds are non-local

In images, similar neighboring pixels can often be assumed to belong to the same visual object but in sound, frequencies are most often non-locally distributed on the spectrogram [4]. Periodic sounds are typically comprised of a fundamental frequency and a number of harmonics which are spaced apart by relationships dictated by the source of the sound. It is the mixture of these harmonics that determines the timbre of the sound.

In the instance of a female vocal, the fundamental frequency at a moment in time might be 200Hz while the first harmonic is 400Hz, the next 600Hz and so on. **These frequencies are not locally grouped but they move together according to a common relationship.** This further complicates the task of finding local features in spectrograms using 2D convolutions as they are often unevenly spaced apart even though they move according to the same factors.



An illustration of the non-local distribution of frequencies in a female voice

Sound is inherently serial

When assessing a visual environment, we can "scan" our surroundings multiple times to locate each visual object in a scene. Since most objects are non-moving, light will reflect from them in a predictable manner and one can make a mental map of their placement in a physical scene. From a perceptual point of view then, the visual objects are assumed to continue to exist at their observed location, even when you look elsewhere.

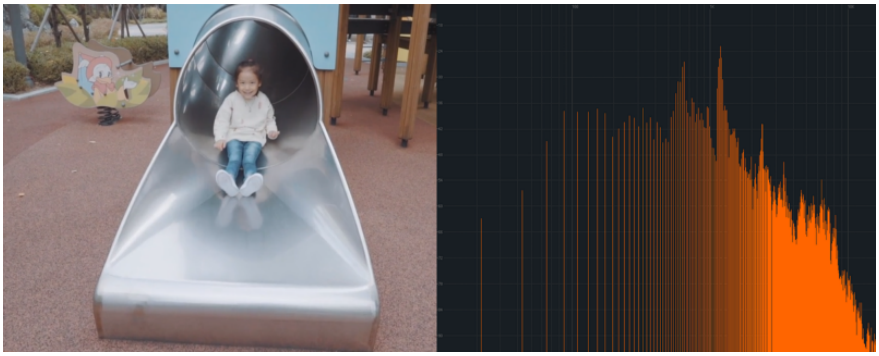
This is not true for sounds. Sound takes the physical form of pressure waves and, from the point of view of a listener, such waves exist only in their current state at one moment in time. Once the moment has passed, the wave has passed by, traveling away from the observer. This is why it makes sense to refer to these phenomena as sound *events* rather than *objects*. From a physical perspective, this means that listeners experience sound only a moment at a time. Where images can be regarded to contain larger amounts of static parallel information, sound, then, is highly serial.

A more fitting comparison is one between audio and video. Both these media can be conceptualized as depicting movements over time where dependencies across time are essential to the experienced meaning of the content. Since video is constructed from collections of images (*frames*), it contains much more parallel information.

One way to illustrate this is to *"freeze"* a moment of time in both media. Looking at a single frame of a video (often depicting ~1/25 seconds of light exposure) it is still often possible to gather a significant amount of meaning about the context, actions and scene of the video: Individual objects can be identified and, sometimes, actions and movements can be assessed. When *"freezing"* a single moment of audio (such as a corresponding aggregate of ~1/25 seconds) with spectral analysis however, assessments cannot be nearly as comprehensive. Some context about the overall tonal balance and characteristics of the signal can be gathered, but not nearly to the same degree as for video.

For example, it is not possible to identify separate sound events outside the context of time to see which spectral developments happen according to the same temporal patterns. The only thing that can be established for certain is the tonal balance of the heard sound(s) at that specific moment in time. An explanation for this comes back to the previously discussed physical form of sound as waves: **Sounds do not**

exist as static objects which can be observed in parallel, they arrive as sequences of air pressure and meaning about these pressures must be established over time.



~1/25 second of video and audio respectively. (Left): A girl riding down a metal slide in a playground. (Right): A spectral representation of a traditional music performance from Senegal.

These reasons suggest that audio as a medium for conveying meaning is fundamentally serial and more temporally dependent than video which presents another reason why visual spectrogram representations of sounds fed into image processing networks without temporal awareness might not work optimally.

A case for modeling the human experience

Significant breakthroughs in AI technology have been achieved through modeling human systems. While artificial neural networks are mathematical models which are only loosely coupled with the way actual human neurons function, their application in solving complex and ambiguous real-world problems has been profound. Modeling the architectural depth of the brain in these neural networks has opened up broad possibilities in learning more meaningful representations of data. In image recognition and processing, the inspiration from the complex and more spatially invariant cells of the visual system in CNNs have also produced great improvements to the state of our technologies.

As argued by J. B. Allen in *"How Do Humans Process and Recognize Speech?"*, as long as human perceptual capacity exceeds that of machines, we stand to gain by understanding the principles of human systems [5]. **Humans are generally very skillful when it comes to perceptual tasks and the contrast between human understanding**

and the status quo of AI becomes particularly apparent in the area of machine hearing. Considering the benefits reaped from getting inspired by human systems in visual processing (and the arguments presented that visual models do not perform as well for sound), I propose that we stand to gain from a similar process in machine hearing with neural networks.

. . .

This is part of a bigger machine hearing project. If you've missed out on the other articles, click below to get up to speed:

Background: The promise of AI in audio processing
Part 1: Human-Like Machine Hearing With AI (1/3)
Part 2: Human-Like Machine Hearing With AI (2/3)

Thanks for reading! To stay in touch, please visit our website neurospace.io and feel free to connect with me on LinkedIn!

References:

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423.

[2] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying Neural Style Transfer," Jan. 2017.

[3] P. Verma and J. O. Smith, "Neural Style Transfer for Audio Spectrograms," Jan. 2018.

[4] L. Wyse. 2017. Audio Spectrogram Representations for Processing with Convolutional Neural Networks. **Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN. Anchorage, US. May, 2017.** 1(1). pp 37–41. DOI: 10.13140/RG.2.2.22227.99364/1

[5] J. B. Allen, "How Do Humans Process and Recognize Speech?," IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 567–577, 1994.