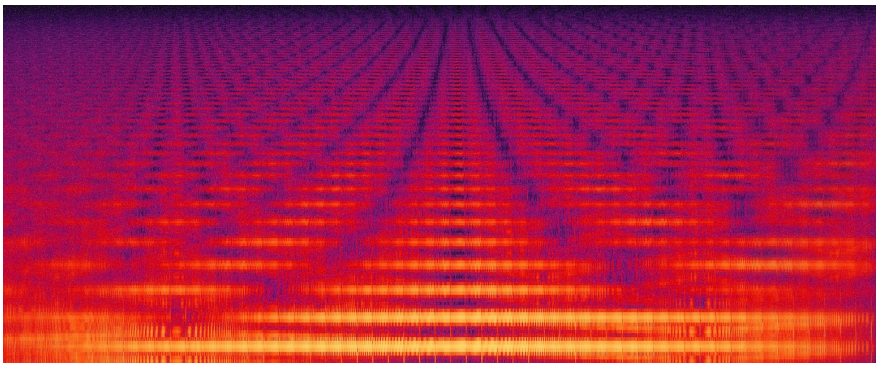
Daniel Rothmann Follow
AI Engineer @ Convai. Especially interested in audio and time series forecasting. Reach us at convai.dk
Dec 26, 2017 · 7 min read



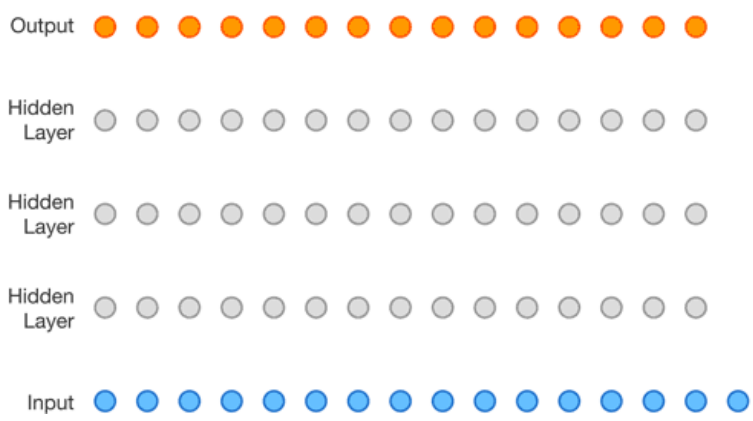
# The promise of AI in audio processing

2017 has been a good year for AI, deep learning in particular. We have seen a rise of AI technologies for image and video processing. Even though things tend to take a little while longer making it to the world of audio, here we have also seen impressive technological advances.

In this article, I will summarize some of these advances, outline further potentials of AI in audio processing as well as describe some of the possible pitfalls and challenges we might encounter in pursuing this cause.

## Towards smarter audio

The kicker for my interest in AI use cases for audio processing was the publication of Google Deepmind's *"WaveNet"*—A deep learning model for generating audio recordings [1] which was released during the end of 2016. Using an adapted network architecture, a *dilated convolutional neural network*, Deepmind researchers succeeded in generating very convincing text-to-speech and some interesting music-like recordings trained from classical piano recordings.




An illustration of WaveNet's dilated model for sample generation (photo credit: Google Deepmind)

In the commercial world, we have also seen more applications of machine learning in products—Take for example LANDR, an automated audio mastering service which relies on AI to set parameters for digital audio processing and refinement.

This year, pro audio software mogul iZotope released Neutron 2, an audio mixing tool that features a *"track assistant"* which utilizes AI to detect instruments and suggest fitting presets to the user. In more direct processing of audio with AI, iZotope also featured a utility for isolating dialogue in their audio restoration suite RX 6.

A short demonstration of iZotope's dialogue isolate feature

## Potentials of AI in digital signal processing

We are still in the early days of AI application in audio processing. Deep learning methods allow us to approach signal processing problems from a new perspective which is still largely ignored in the audio

industry. Up to this point, we have had our focus on formulaic treatment: Coming to a deep understanding of a problem and manually devising functions to solve it. However, the understanding of sound is a very complex task and problems that we, as humans, intuitively find quite easy often turn out to be very difficult to describe formulaically.

Take, for instance, source separation: You find yourself in a scenario where two people are speaking over each other. After the fact, in your mind, you can imagine either person speaking in isolation without much effort. But how do we describe a formula for separating these two voices? Well, it depends:

> *Is there a unified way to describe how human voices sound? If yes, how are parameters of this description affected by sex, age, energy, personality? How does physical proximity to the listener and room acoustics impact this understanding? What about non-human noise that can occur during the recording? On which parameters can we discriminate one voice over another?*

As you can see, devising a formula for the full extent of this problem would require attention to a lot of parameters. Here, AI can provide a more pragmatic approach—By setting up the proper conditions for learning, we can statistically estimate the complexities of this function automatically. In fact, researchers at Eriksholm (a research center for the hearing aid manufacturer Oticon) recently proposed a method for achieving improved source separation in real-time applications using a convolutional recurrent neural network architecture [2].

As methods for processing audio with deep neural networks are improving, we can only begin to imagine the difficult problems we could solve—Here are a few of my imaginations for deep learning in real-time audio processing:

- **Selective noise canceling**, removing only certain elements like car traffic
- **Hi-fi audio reconstruction**, from small and low-quality microphones
- **Analog audio emulation**, estimating the complex interactions between non-linear analog audio components

- **Speech processing**, changing speaker, dialect or language in recordings
- **Improved spatial simulations**, for reverb and binaural processing

## Challenges in representation and architecture

WaveNet was among the first successful attempts to generate audio on a raw sample level. The one big problem here is that CD quality audio is usually stored with 44100 samples per second and thus, generating *seconds* of sound with WaveNet takes *hours*. This excludes the method from having use in real-time applications. It's just a lot of data to make sense of.

On the other hand, many current solutions for audio processing with neural networks utilize spectrogram representations and convolutional networks. In these solutions, the audio frequency spectrum is essentially represented visually as magnitudes over time on a 2D image and the convolutional network is used to scan and process that image [3]. Usually, results from these methods are not nearly as compelling as those we see in the visual field such as *CycleGAN* that can do some impressive style transfer for movies [4].

CycleGAN transforming horses into zebras (photo credit: CycleGAN)

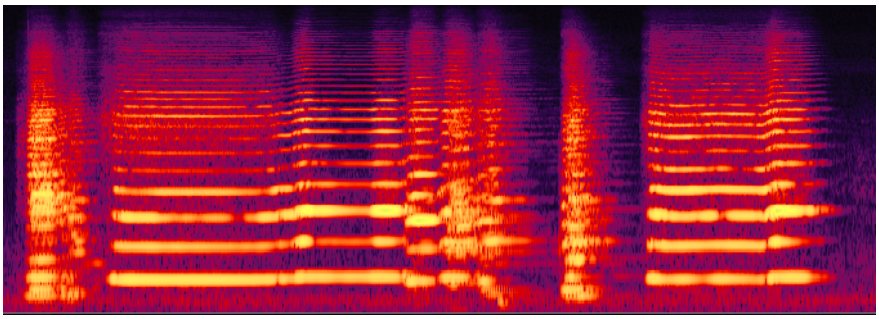
Movies and audio clips have something in common in the sense that they both depict movements over time. Considering innovations in image processing networks like CycleGAN, one would assume that such style transfer could be possible for audio as well.

But movies and audio clips are not the same things—If you freeze a frame of a movie, quite a lot of information can still be gathered about the actions in the frame. If you freeze a "frame" of audio, however, very little meaning can be gathered about what is actually being heard. This suggests that audio is fundamentally more temporally dependent than

movies. In spectrograms, you can never assume that a pixel belongs to a single object either: Audio is always “transparent”, and therefore spectrograms display all audible sounds overlapping each other in the same frame [3].

*We are doing machine vision to do machine hearing.*

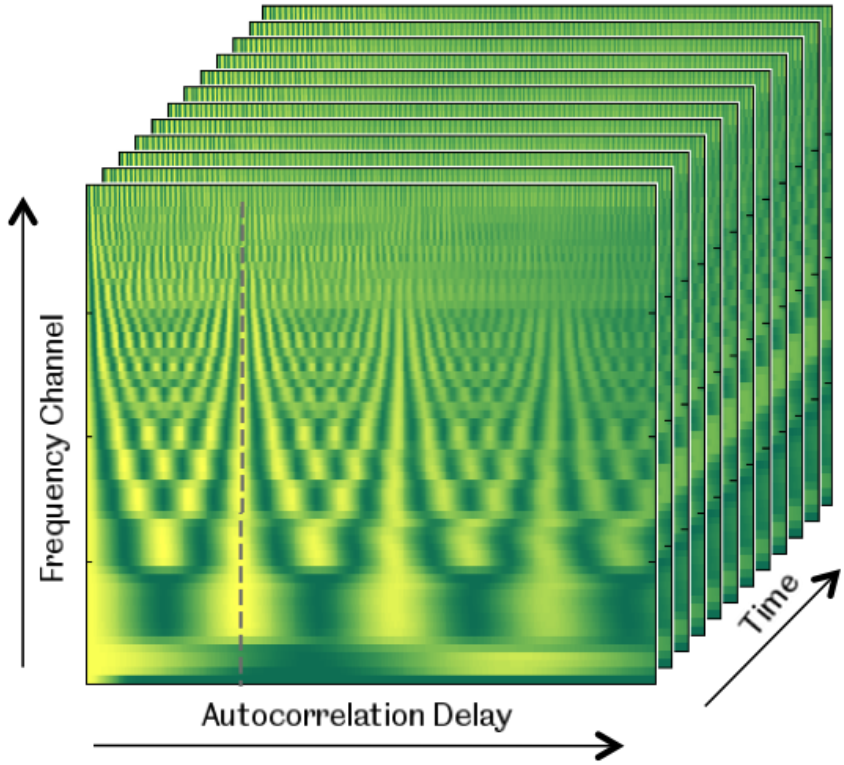
Convolutional neural networks are designed with inspiration from the human visual system, loosely based on how information flows into the visual cortex [5]. I believe this presents a problem worth paying consideration to. Essentially, we are taking audio, translating it to images and performing visual processing on that image before translating it back to audio. So, we are doing machine vision to do machine hearing. But, as we are intuitively aware, these two senses don’t function in the same way. Looking at the spectrogram below, how much meaning can you (with your smart human brain) actually gather about the contents of the audio? If you could listen to it, you would quickly get an intuitive understanding of what is happening. Maybe this is a problem that is hindering our progress in AI-assisted technologies for audio.



A five-second spectrogram. Can you tell what it is? (It’s a blues harp.)

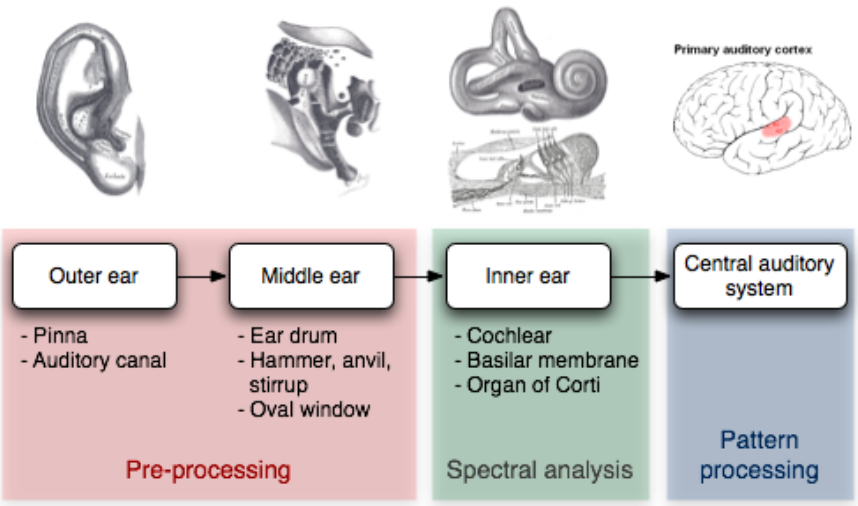
Therefore I propose that to achieve better results with neural networks for audio processing, we should allocate energy to figuring out better representations and neural network architectures for audio specifically. One such representation could be the autocorrelogram, a three-dimensional representation of sound where time, frequency and periodicity is included [6]. It turns out that humans can separate sound sources by intuitively comparing the periodicity of sounds to find patterns of similarity. Pitch and rhythm are also results of temporal

factors. Therefore more temporally focused representations, such as the autocorrelogram, might be useful.




An illustration of an autocorrelogram for sound (photo credit: University of Sheffield)

Furthermore, we could start thinking about architecturally modeling the neural pathways of the auditory system. When sound excites our eardrums and travels into the cochlea, it is transformed into magnitudes for different frequencies. Then it travels to the central auditory system for temporal pattern processing. Which modes of analysis do we utilize to gather meaning from audio in our central auditory system that can be modeled in an artificial neural network? Periodicity, maybe [6]. Statistical groupings of sound events, maybe [7]. Dilated time frames of analysis, maybe [1].




An illustration of information flow in the auditory system (photo credit: Universität Zu Lübeck)

## Conclusion

Developments in AI presents a big potential for smarter audio signal processing. But to better understand sound in neural networks, we might need to part ways with perspectives that are inherently visual and consider new techniques based instead on the auditory system.

In this article, I have posed more questions than I have provided answers, hoping to kick-start your thought about sound in this context.

I am a Master's student at Aarhus University, currently writing my Masters' Thesis on the subject of real-time audio processing with artificial neural networks, for which this article is a sort of primer. My hope is to continually publish articles like this as the project progresses forward.

. . .

Do you have questions or comments? Leave a note or chat me on LinkedIn—I am always open to talk!

## Sources

[1] *A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu*: **WaveNet: A Generative Model for Raw Audio**, 2016

[2] *G. Naithani, T. Barker, G. Parascandolo, L. Bramsløw, N. Pontoppidan, T- Virtanan*: **Low Latency Sound Source Separation Using Convolutional Recurrent Neural Networks**, 2017

[3] *L. Wyse*: **Audio spectrogram representations for processing with Convolutional Neural Networks**, 2017

[4] *J. Zhu, T. Park, P. Isola, A. Efros*: **Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks**, 2017

[5] *Y. Bengio*: **Learning Deep Architectures for AI** (p. 44), 2009

[6] *M. Slaney, R. Lyon*: **On the importance of time—A temporal representation of sound**, 1993

[7] *E. Piazza, T. Sweeny, D. Wessel, M. Silver, D. Whitney*: **Humans Use Summary Statistics to Perceive Auditory Sequences**, 2013