IMPROVED SOUND REPRESENTATION FOR MACHINE LEARNING

Carlos Tarjano

PhD student at UFF – RJ, Brazil

tesserato@hotmail.com

Github

SoundCloud

ABOUT CARLOS TARJANO

- Works at UFRJ
- Bachelor (CEFET/RJ) and Master's degree(UFF) in Industrial Engineering
- Amateur musician – Drums and acoustic guitar

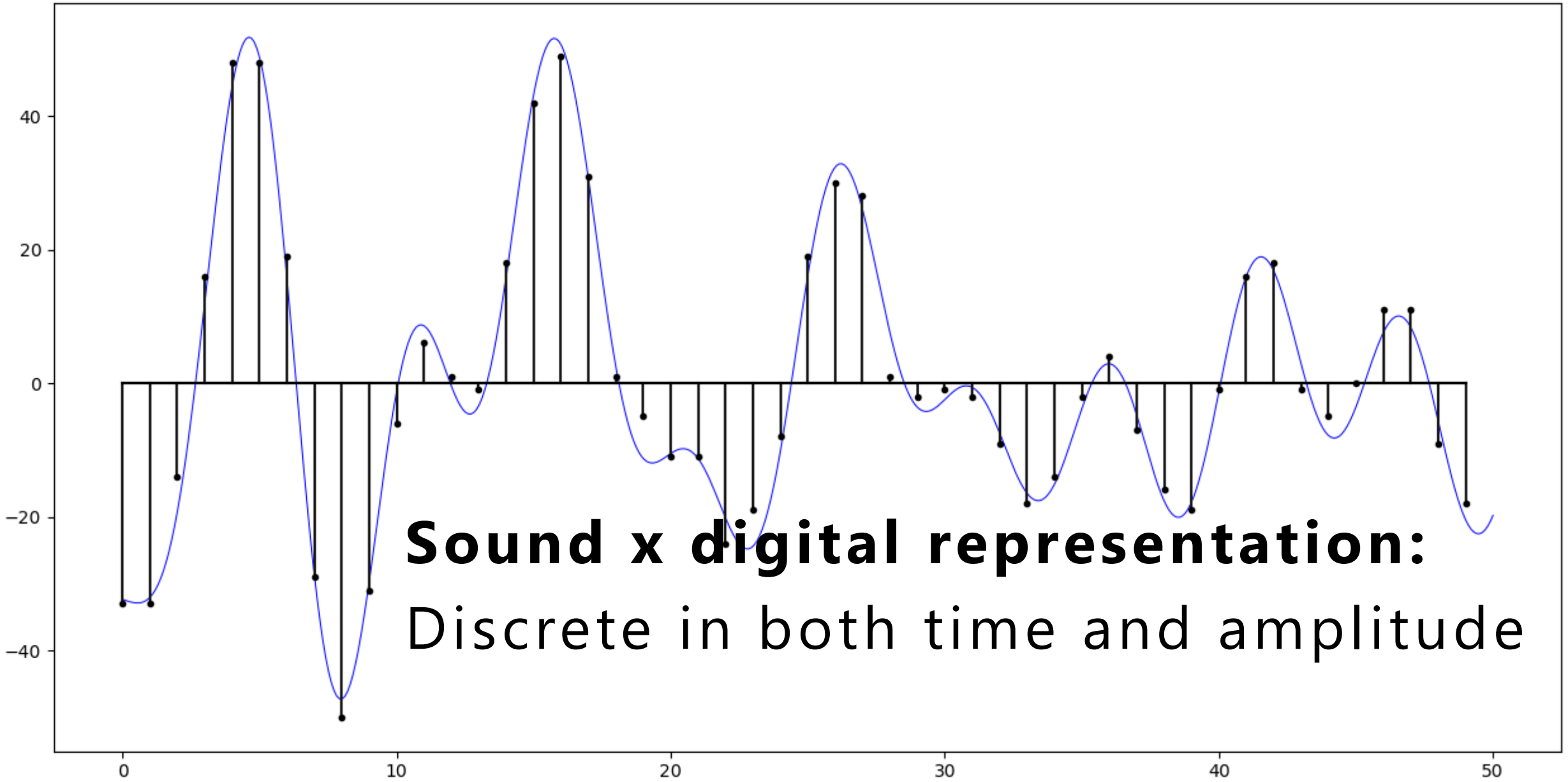
RESEARCH AND INTERESTS

- Digital Musical Instruments
- Product development background
- Neural Networks
- Programming

STEP BACK – WHAT IS SOUND

- Fluctuations in air pressure
- We can process them from 20 hz to 20 kHz
- Some sounds are more “pleasant” than others

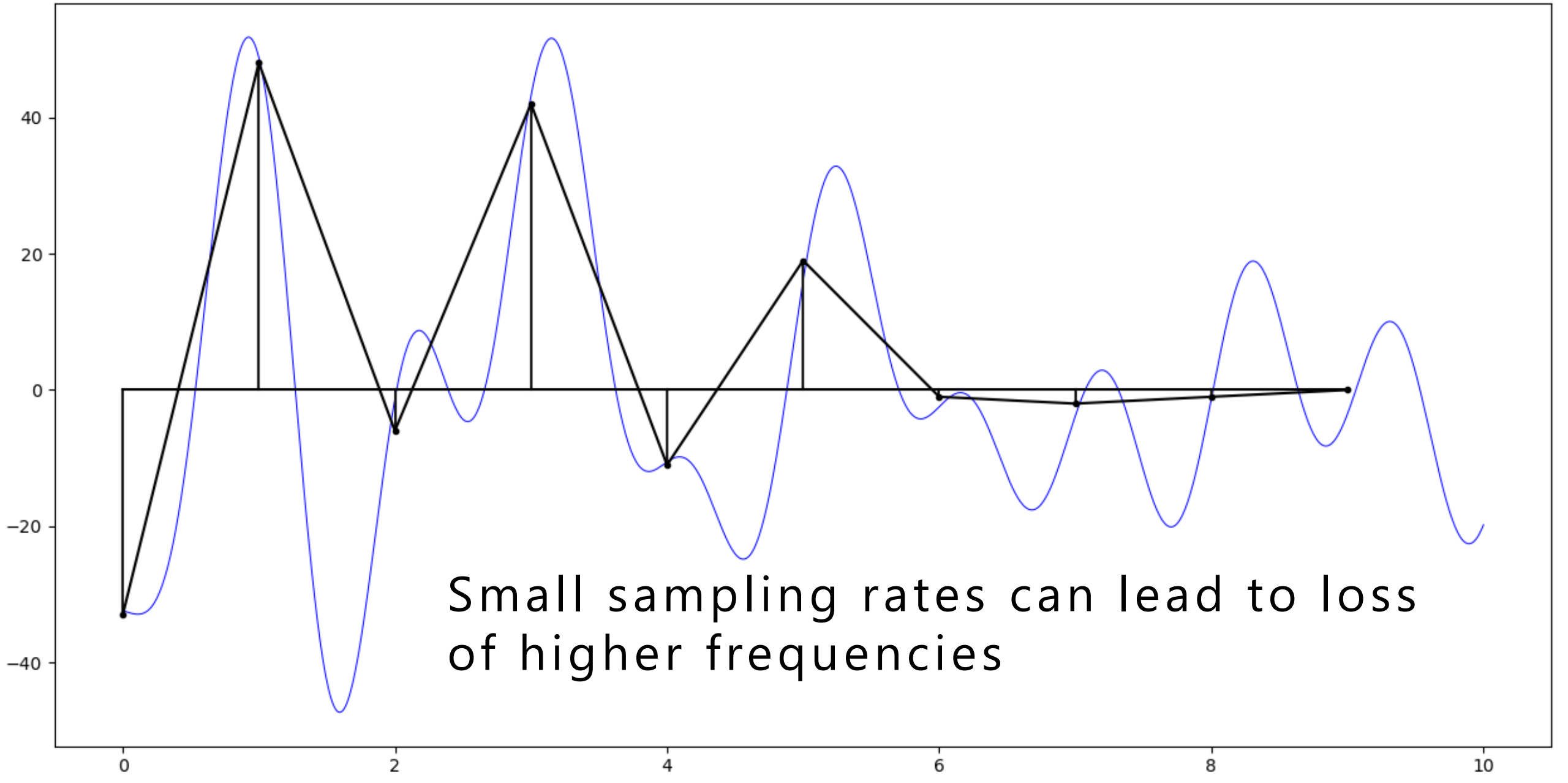
DIGITAL REPRESENTATION – TIME DOMAIN



Sound x digital representation:

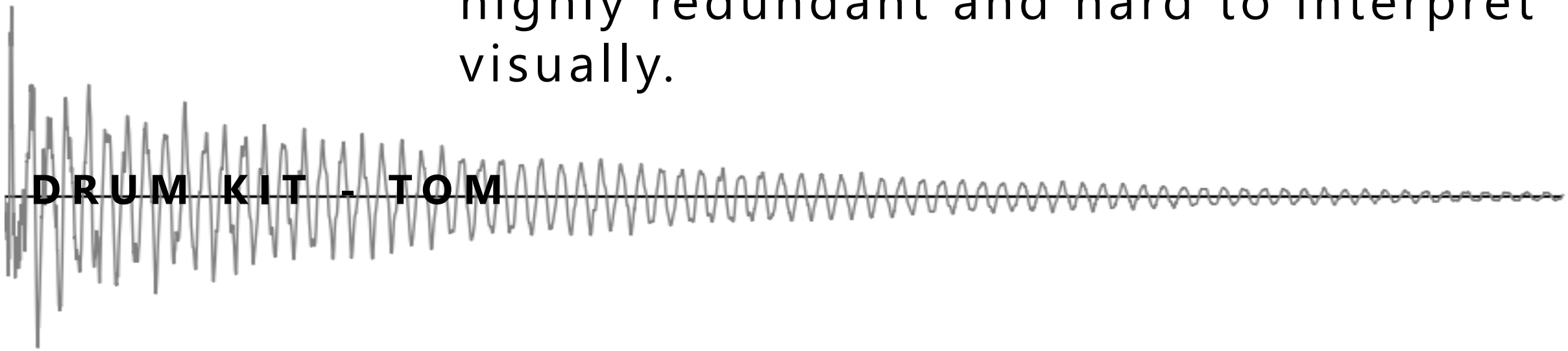
Discrete in both time and amplitude

DIGITAL REPRESENTATION – TIME DOMAIN



DIGITAL REPRESENTATION – TIME DOMAIN

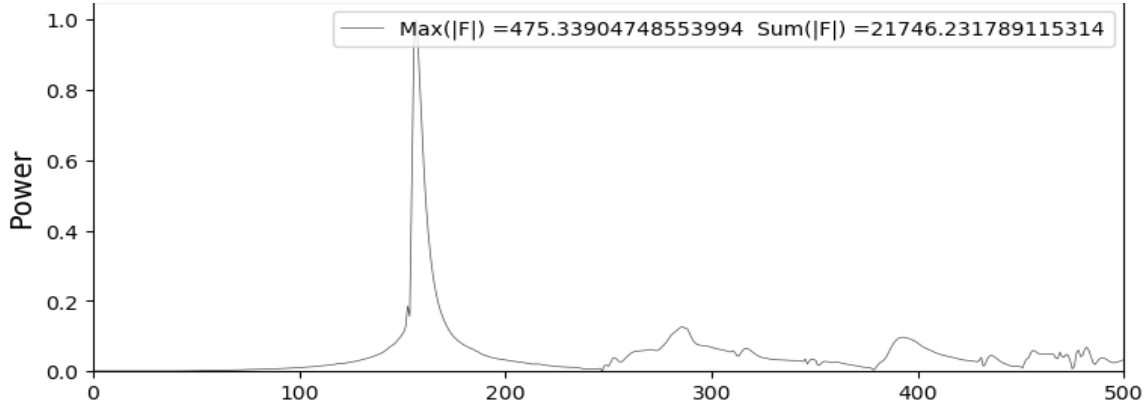
Time domain representations are straightforward and accurate, but highly redundant and hard to interpret visually.

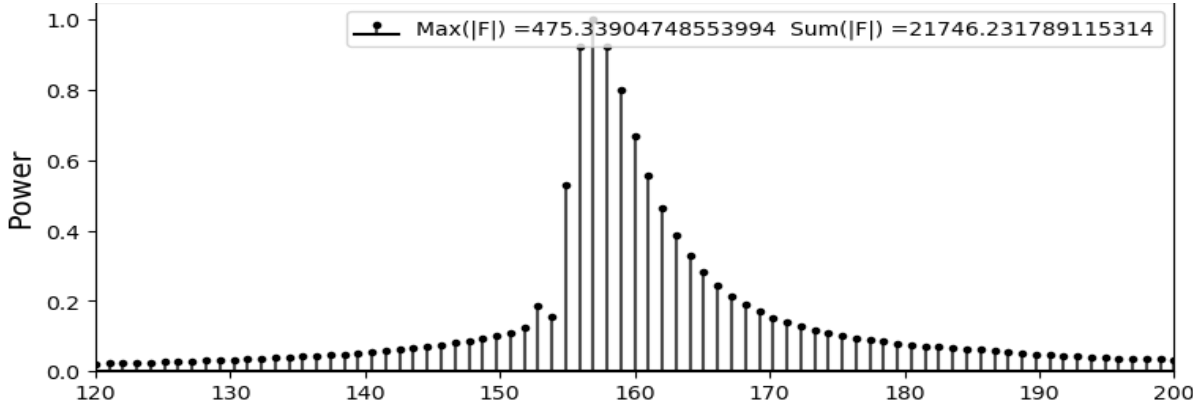


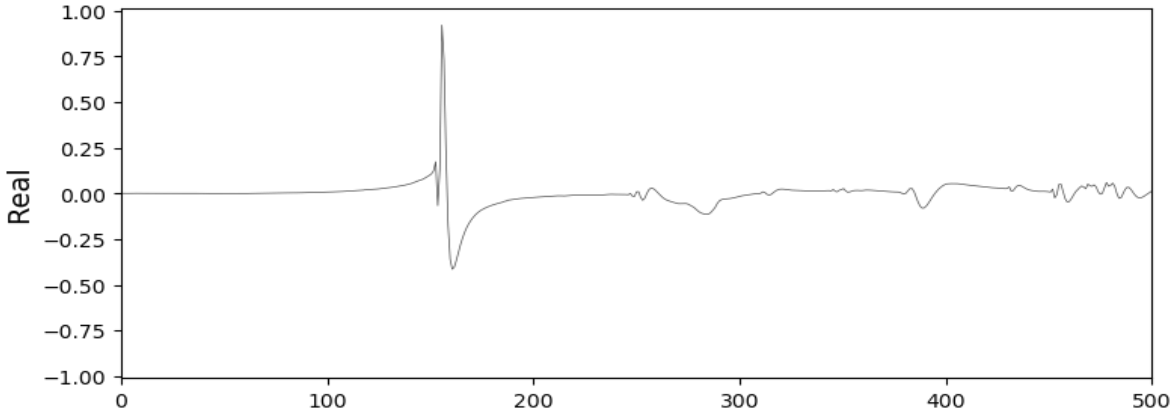
DIGITAL REPRESENTATION - FREQUENCY DOMAIN-FT

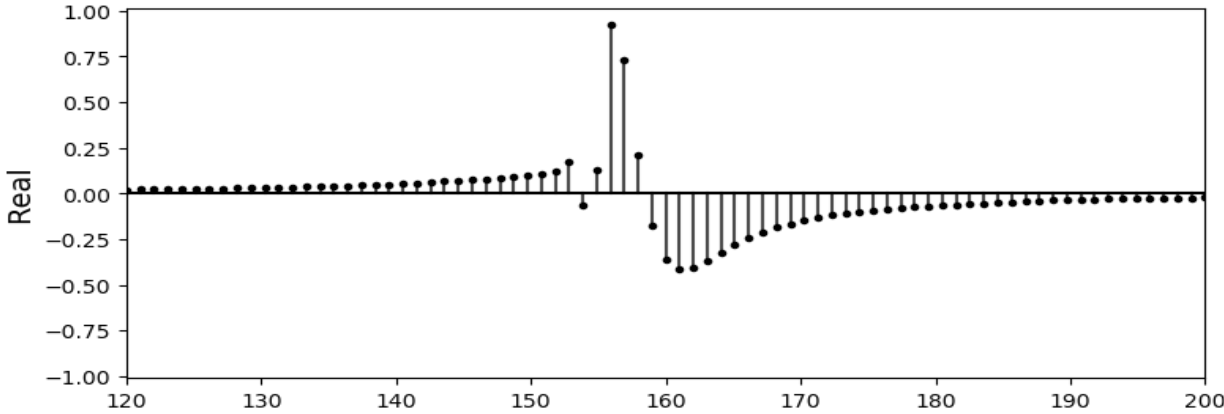
- PROS:
- (Discrete) Fourier Transform – “natural”
- Revertible
- Relatively fast algorithms

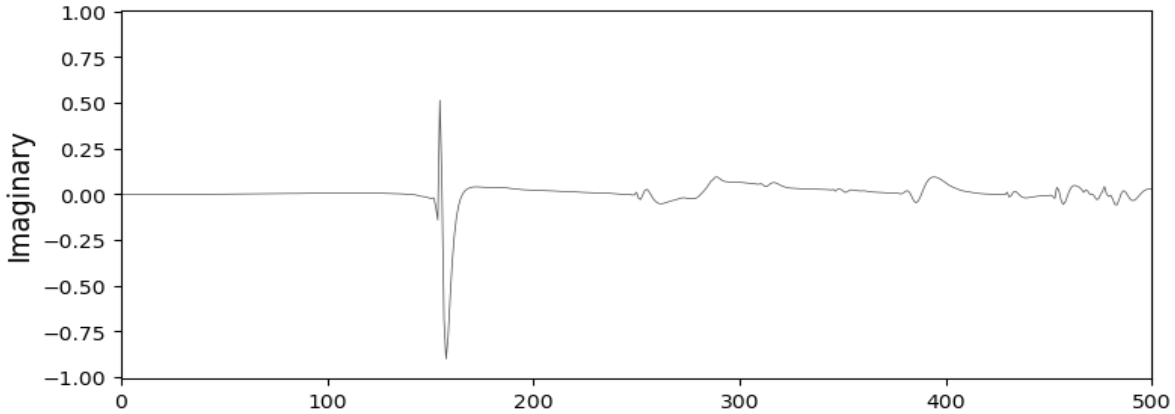
DIGITAL REPRESENTATION - FREQUENCY DOMAIN-FT

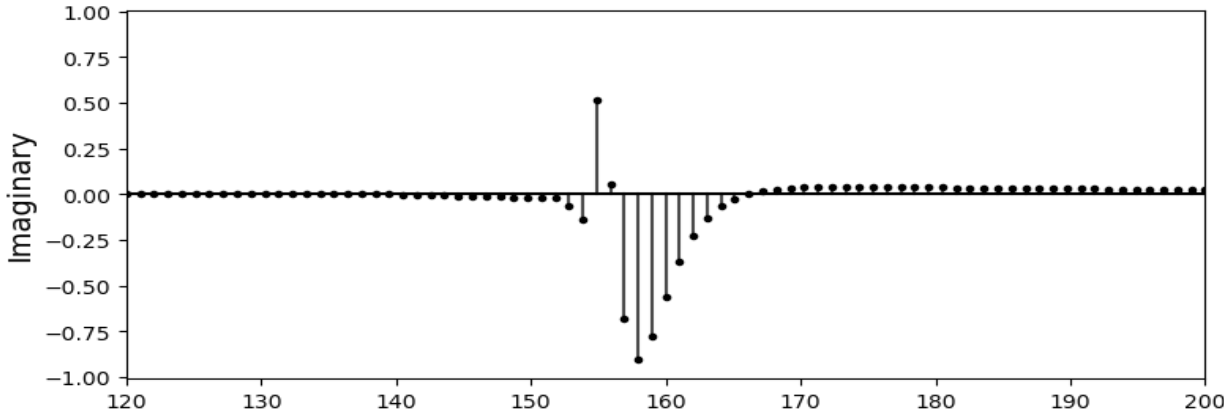




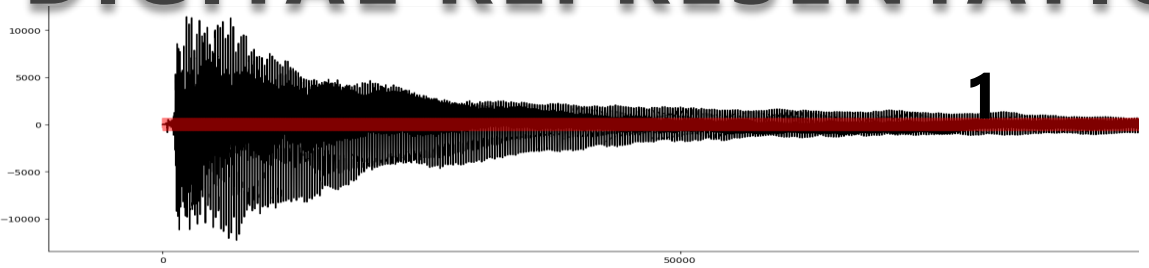


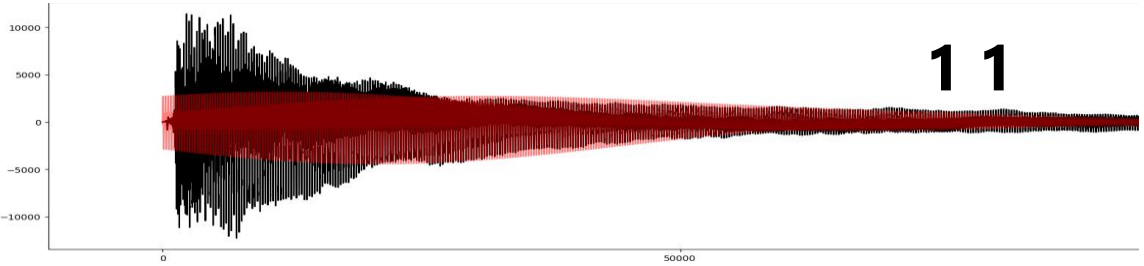


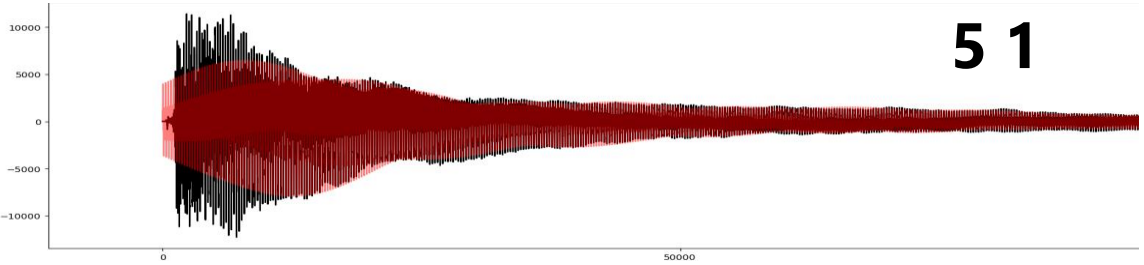


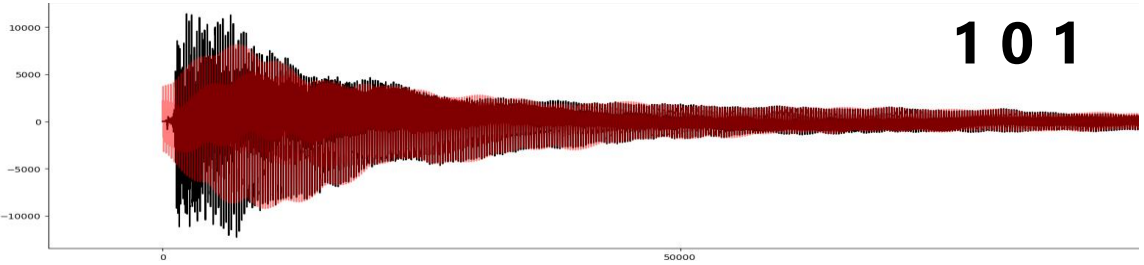


DIGITAL REPRESENTATION - FREQUENCY DOMAIN-FT







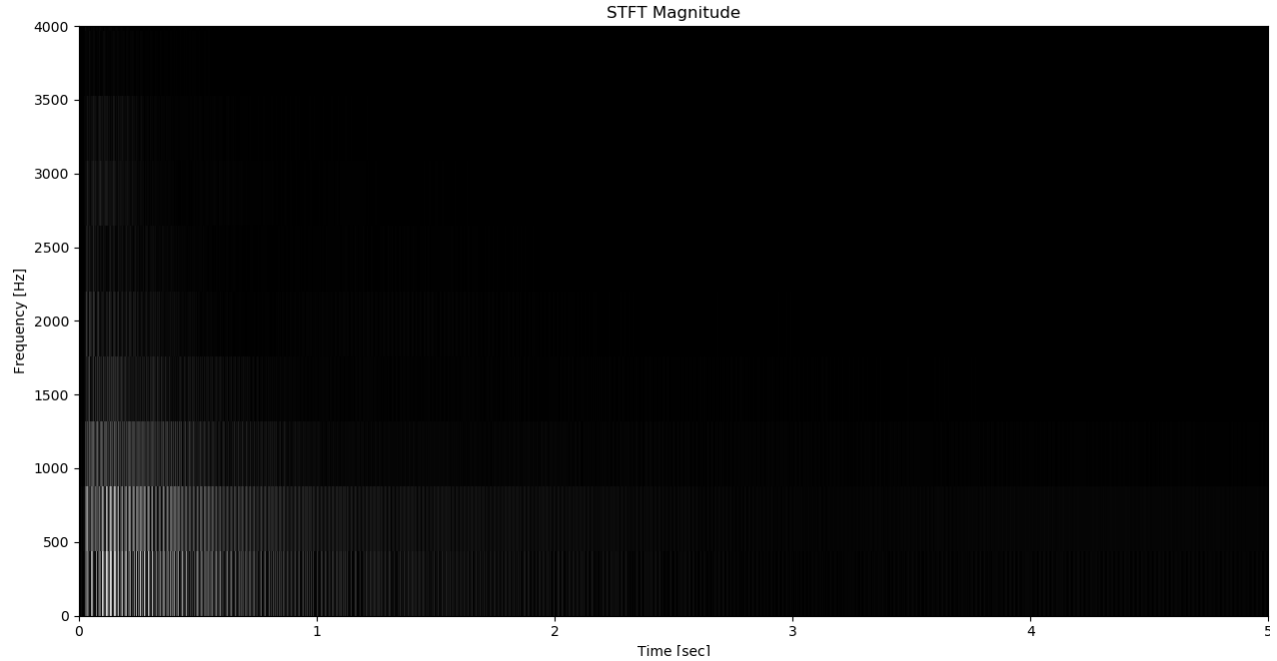


We can see to the left a complex wave being represented by an increasing number of sinusoids, extracted from the DFT. One can note that the DFT is very conservative: a vast number of sinusoids are needed to approximate the original signal to a reasonable degree. Also, the returns are diminishing: adding 10 sinusoids from 1 to 11 improves the representation much more than adding 50, from 51 to 101.

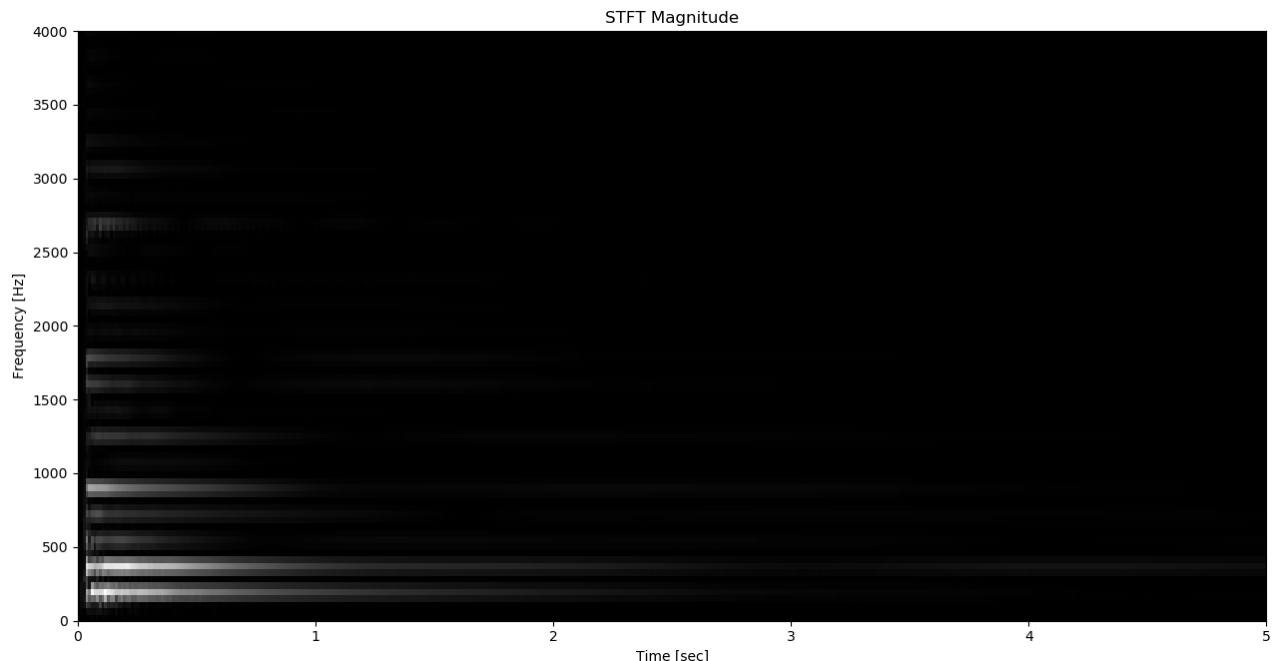
DIGITAL REPRESENTATION - MIXED

- DFTs have no time localization. The Short-time Fourier Transform approaches this problem by dividing the signal in smaller segment and applying DFT in each one; There is a trade off between time and frequency resolutions.

Window =100 samples

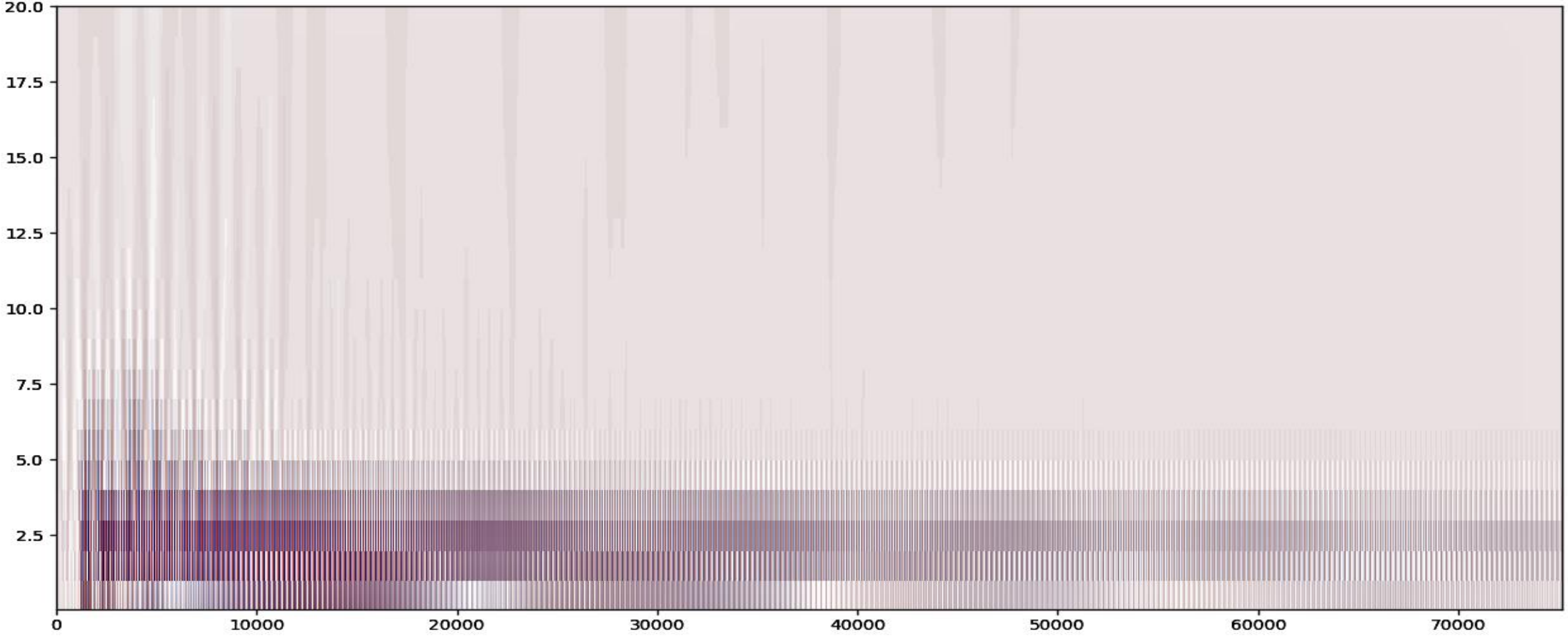


Window =1000 samples



DIGITAL REPRESENTATION - MIXED

- The wavelet transform addresses this problem introducing varying resolutions. Computational cost also increases



DIGITAL MUSICAL INSTRUMENTS

CONTROLLERS

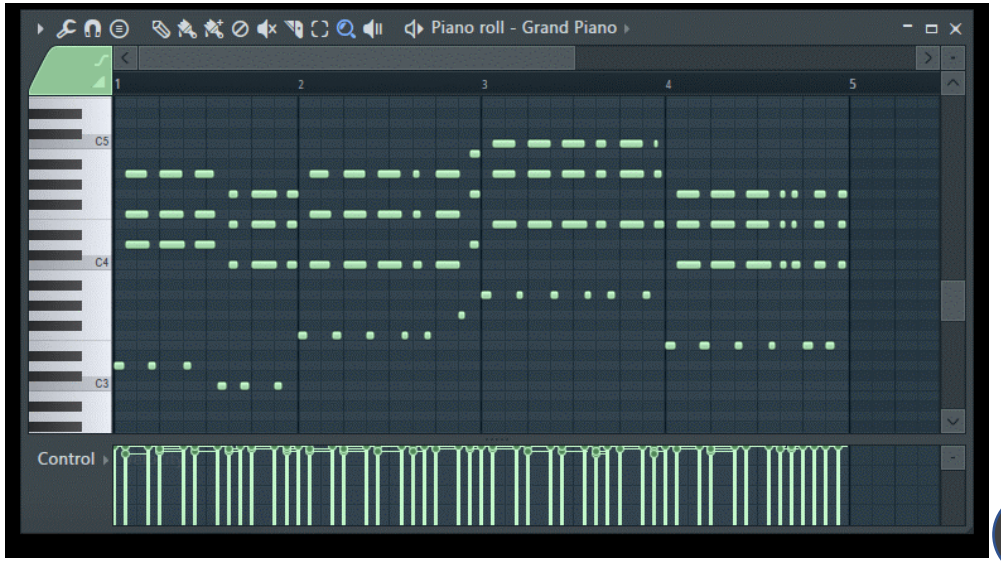
1

2

3

4

NOTATION / MESSAGE



5

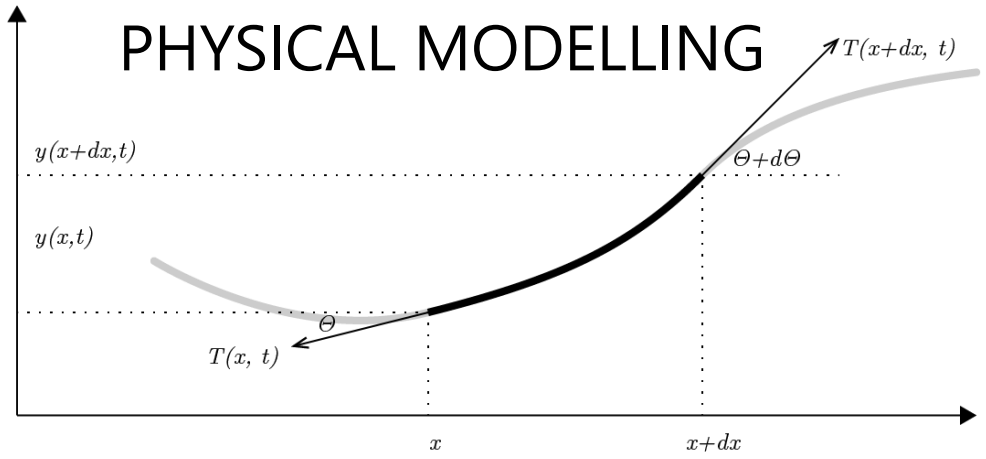
ENGINE

EMULATING REAL WORLD INSTRUMENTS

SAMPLE BASED

- Those are the main approaches used to model real world instruments, with sample based approaches being the industry standard

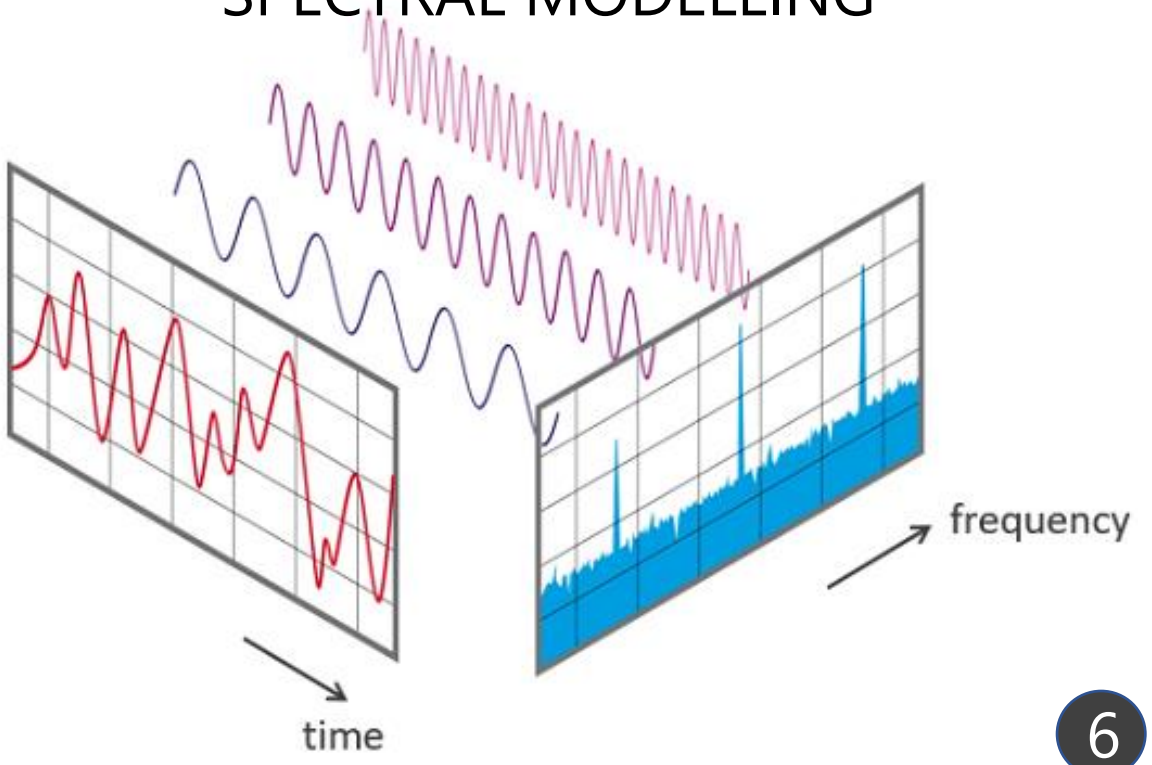
PHYSICAL MODELLING



$$\frac{\partial^2 y(x,t)}{\partial t^2} = c^2 \frac{\partial^2 y(x,t)}{\partial x^2}$$

$$y(x,t) = f(x - ct) + g(x + ct)$$

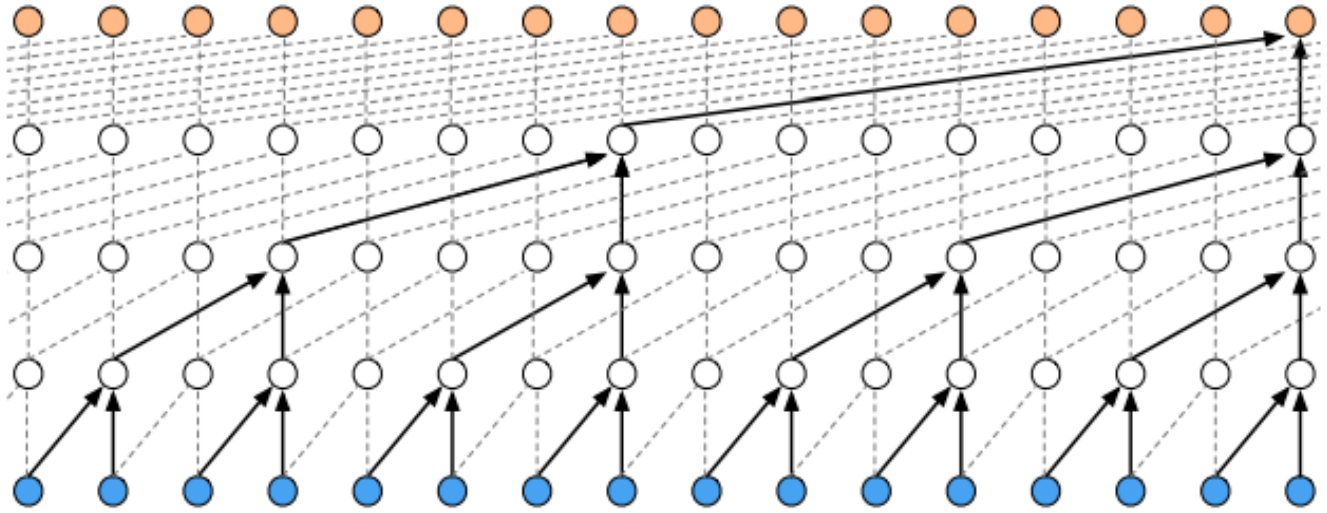
SPECTRAL MODELLING



EMERGING APPROACHES

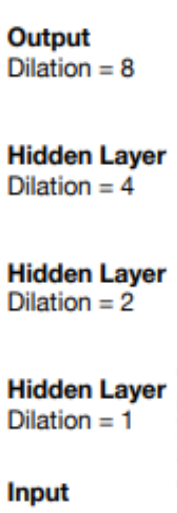
- Machine learning techniques are emerging, for sound generation and classification. Sound generation is still based in time domain representations. More appropriate representations would facilitate the translation of computer vision approaches, bringing a great potential of improvement to the field.

7 Wavenet

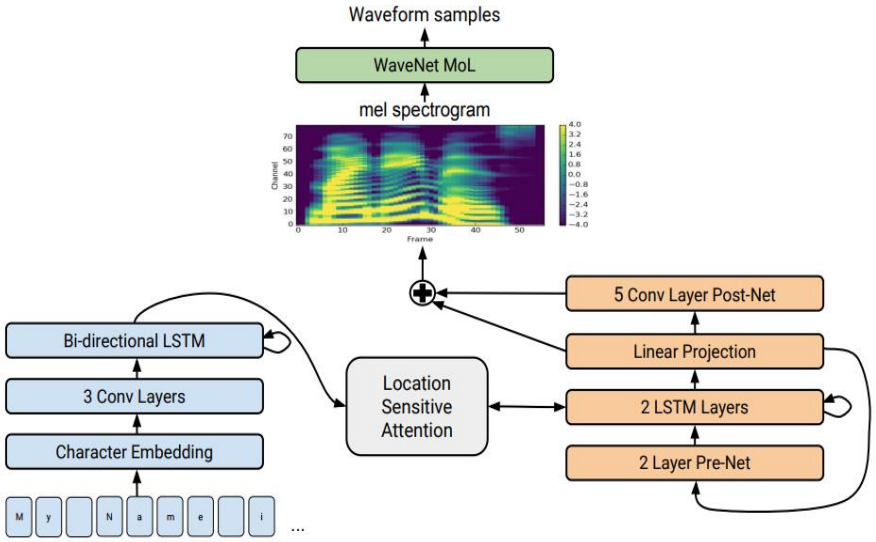


8 DeepVoice I

9 Magenta

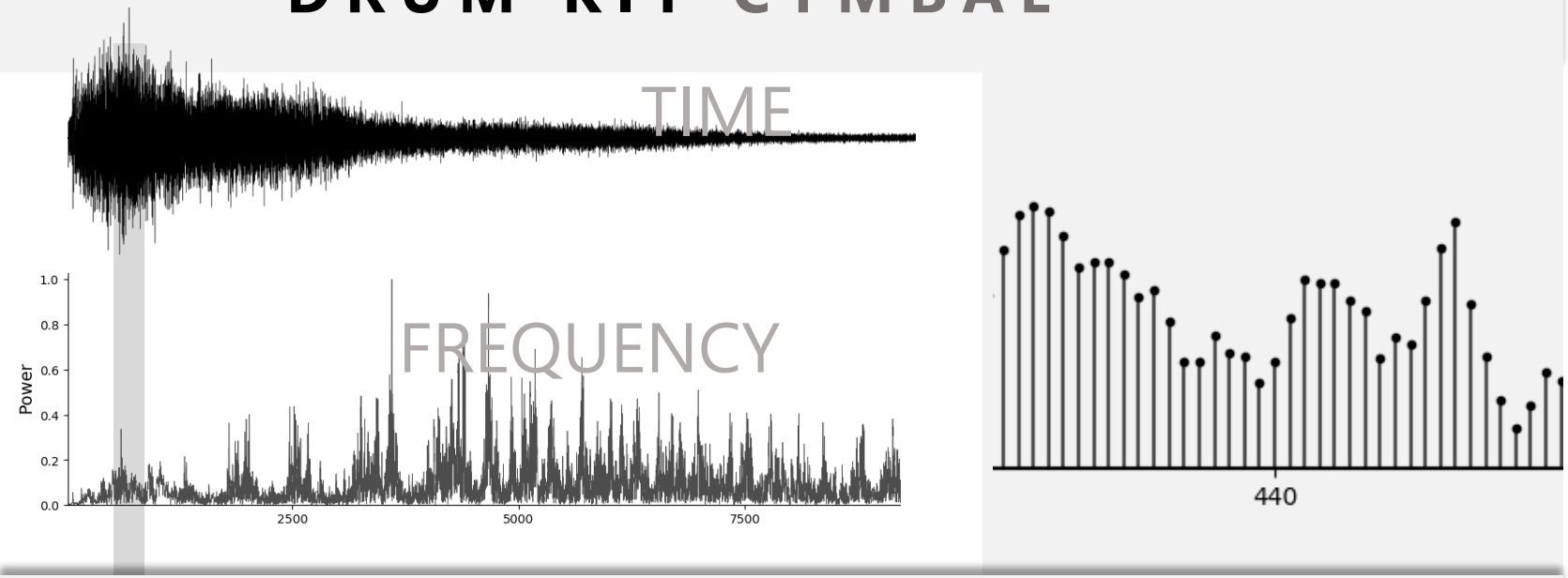


10 Tacotron

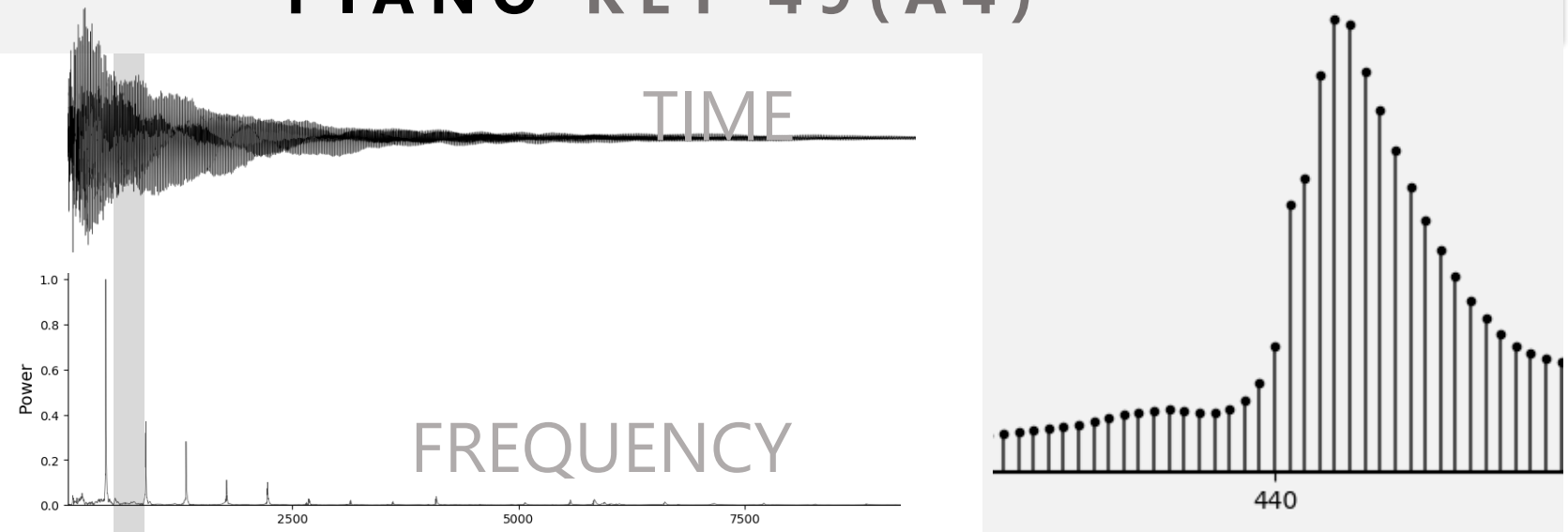


HARMONIC SOUNDS IN THE FREQUENCY DOMAIN

DRUM KIT CYMBAL

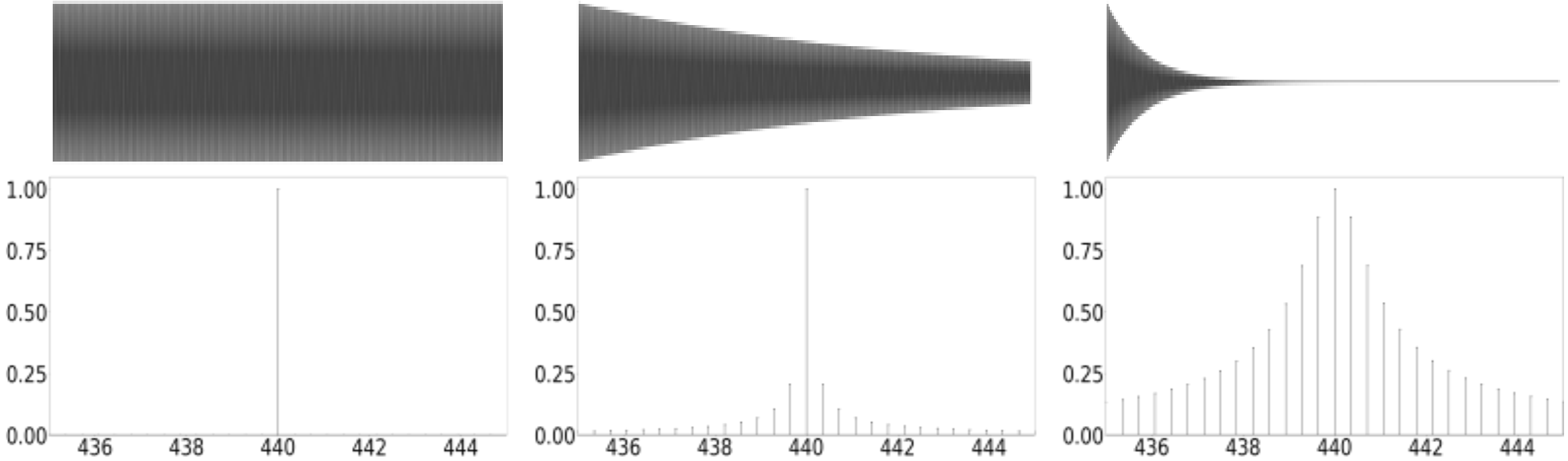


PIANO KEY 49(A4)



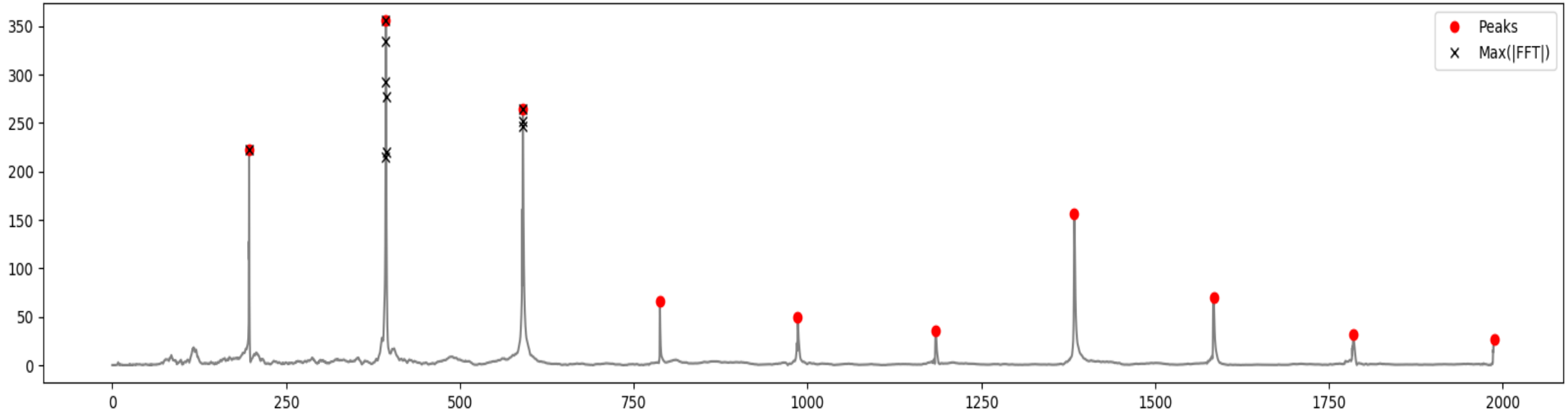
- Harmonic sounds have a well behaved structure in the frequency domain that can be exploited

ENVELOPE INFLUENCE ON FREQUENCIES



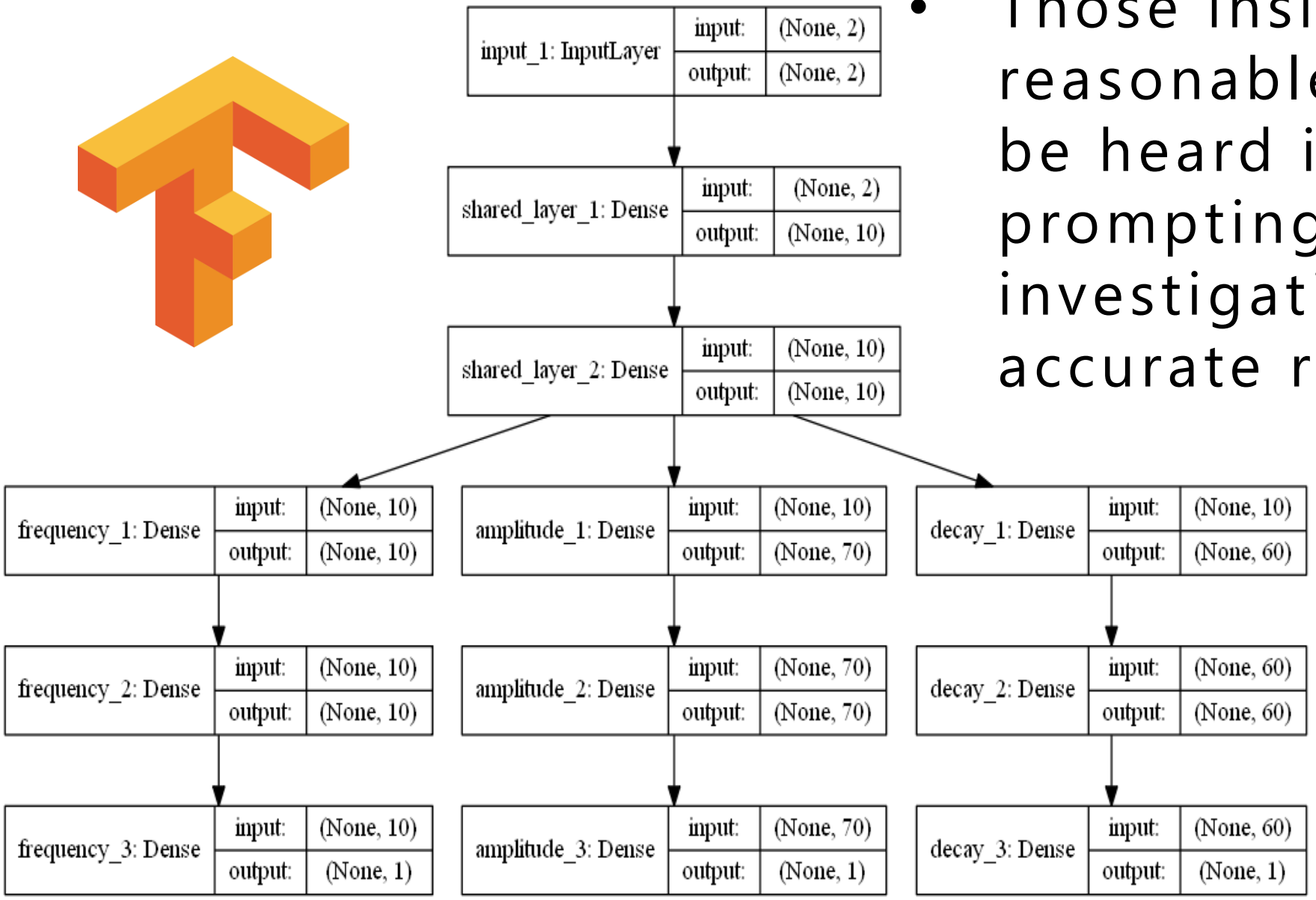
- The image above illustrates the influence of the envelope of the signal and the frequency content in the frequency domain

PEAK DETECTION



- Peak detection is not straightforward in the frequency domain: the figure illustrates peaks in red versus the highest values of the DFT. It requires underlying information of the signal. Still, it illustrates possibilities of alternative representations:

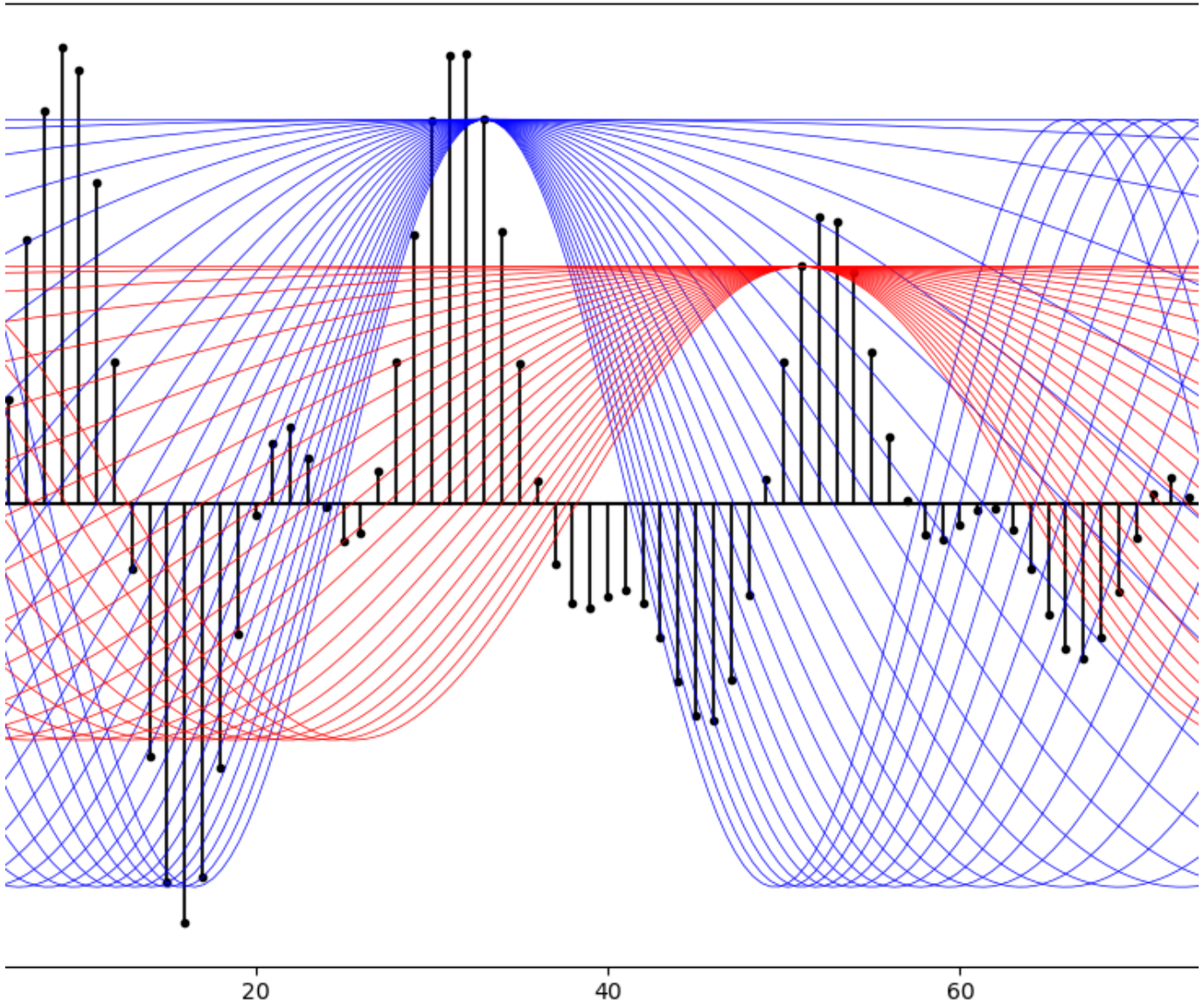
PRELIMINARY RESULTS



- Those insights led to reasonable results, as can be heard in the link below, prompting more investigation towards more accurate representations

SOUNDCLOUD

PROPOSED NEW TRANSFORM



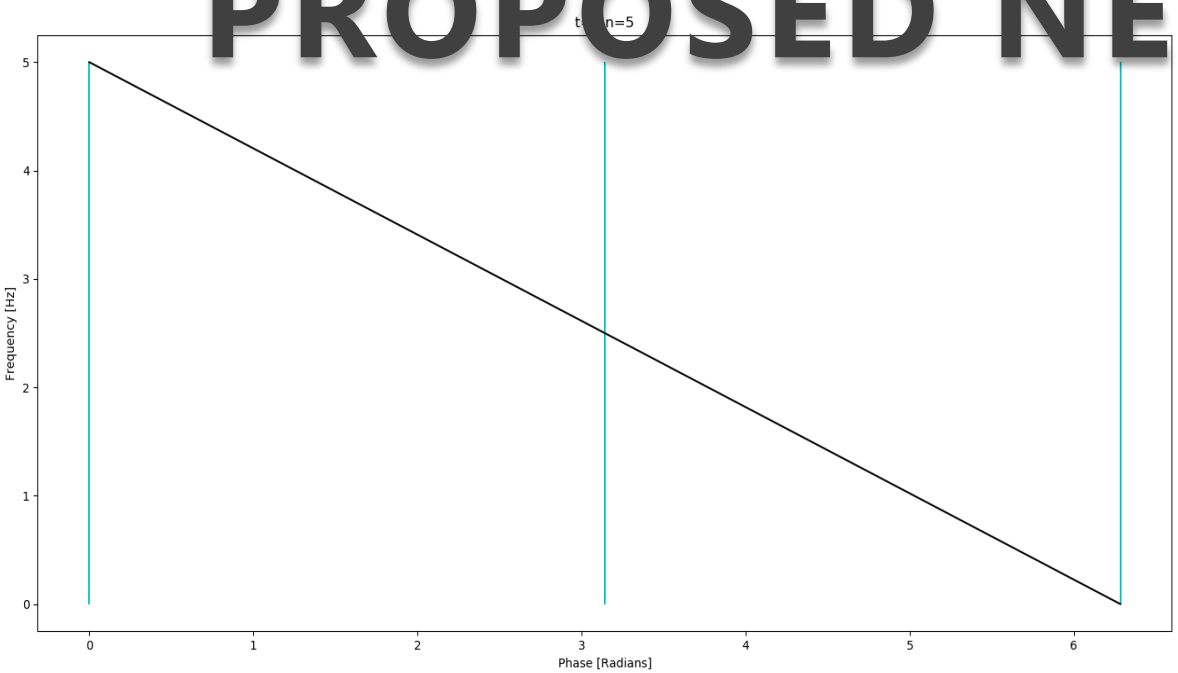
- Some members of the family of sinusoids that have their maximum coinciding with the points in a signal are illustrated in the picture, for two samples

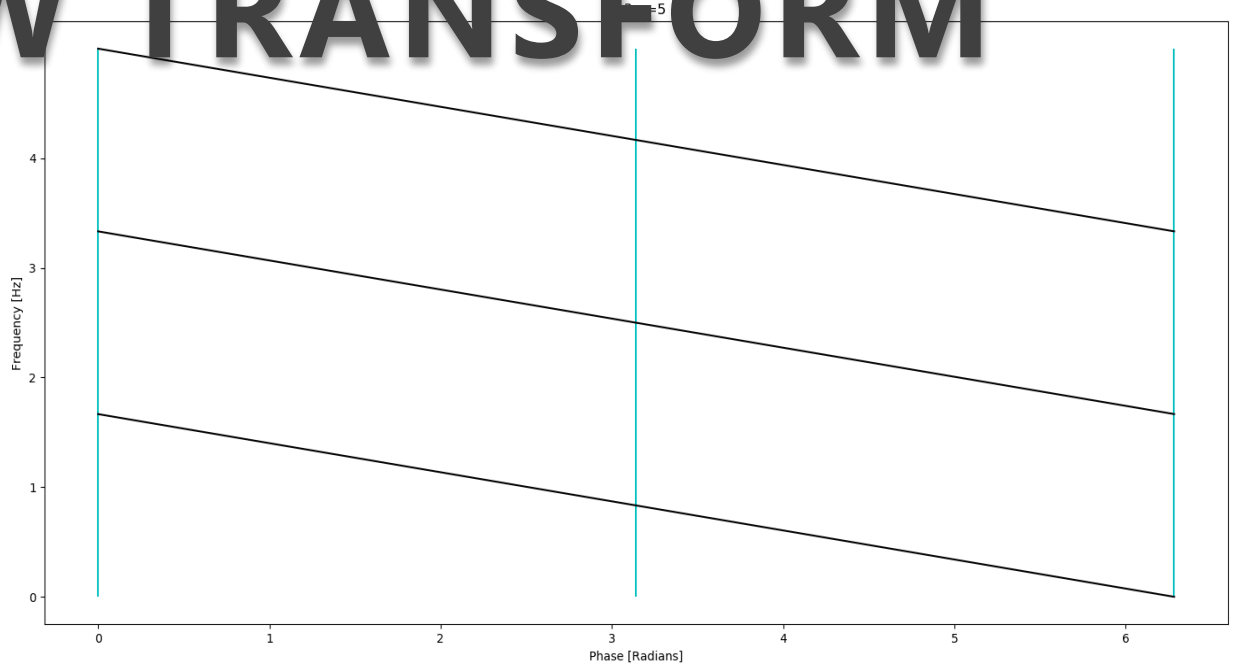
PROPOSED NEW TRANSFORM

$$a\cos\left(\theta+\frac{2\pi f t}{n}\right)=W_t \rightarrow f_{max}=\begin{cases}\frac{nk}{t}-\frac{n\pi}{2\pi t}\ if\ W_t>0\\ \frac{nk}{t}-\frac{n\pi}{2\pi t}+\frac{n}{2t}\ if\ W_t<0\end{cases}$$

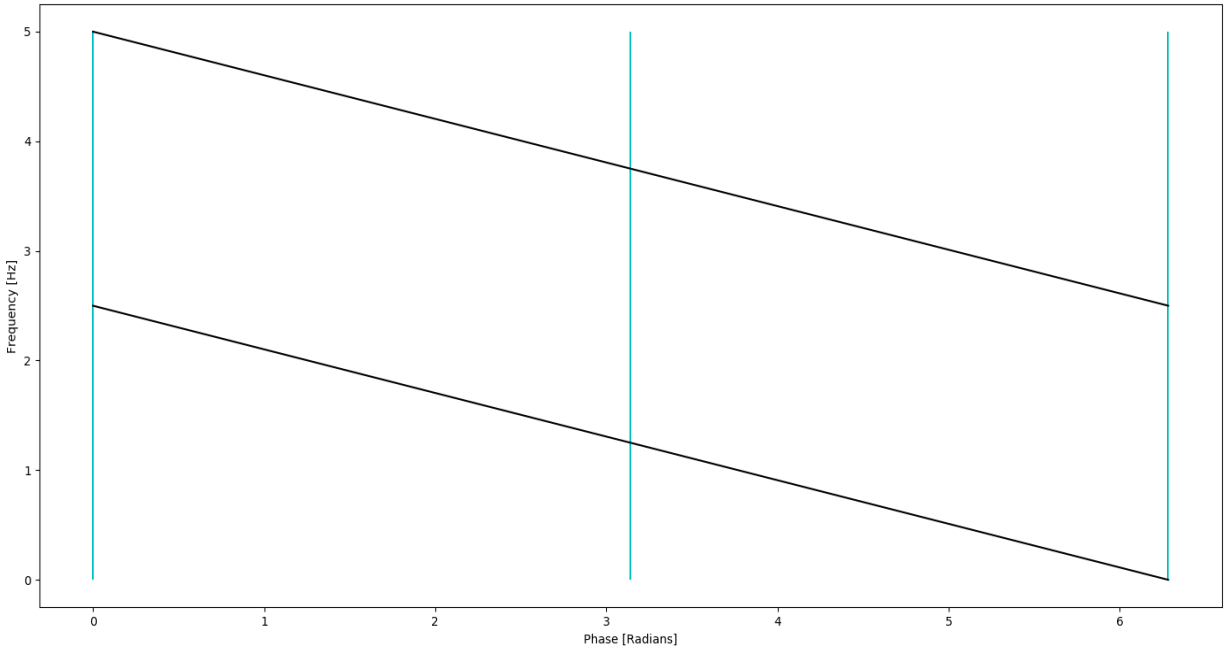
- If we assume that the sinusoids are described by the equation on the left hand side, the maxima are described, as shown in the right hand side, for the two possible cases where the sample is positive and negative, by linear equations. In the following slide, those equations are plotted in the space of possible phases and frequencies.

PROPOSED NEW TRANSFORM

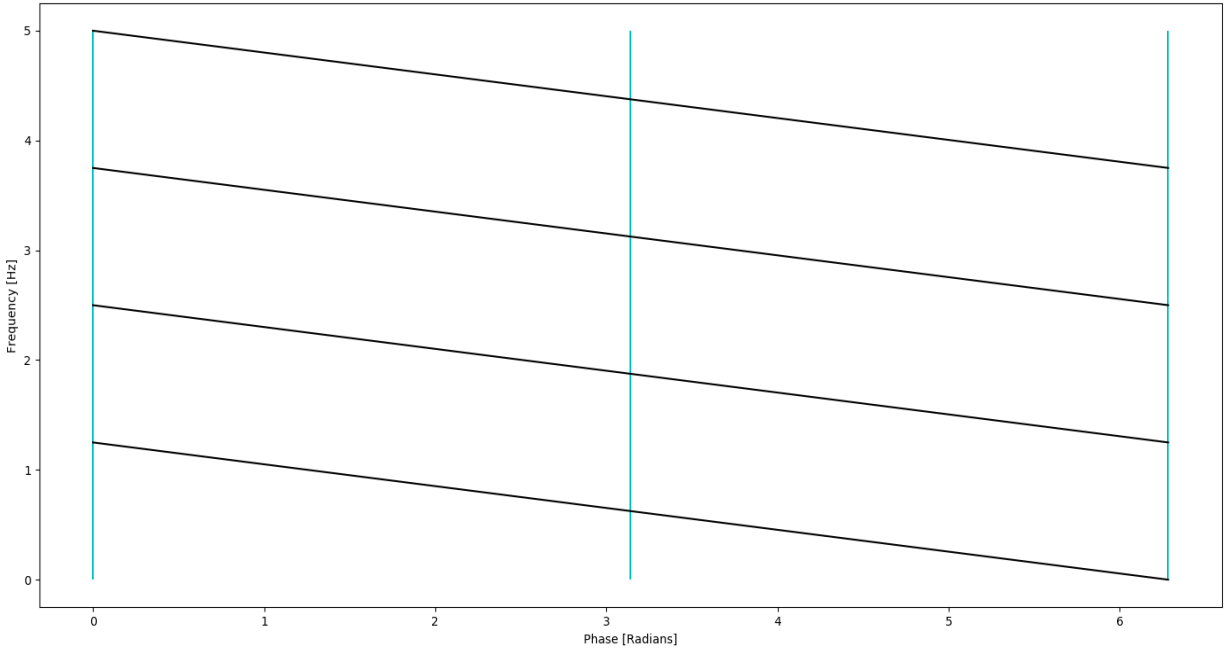




t=2 n=5



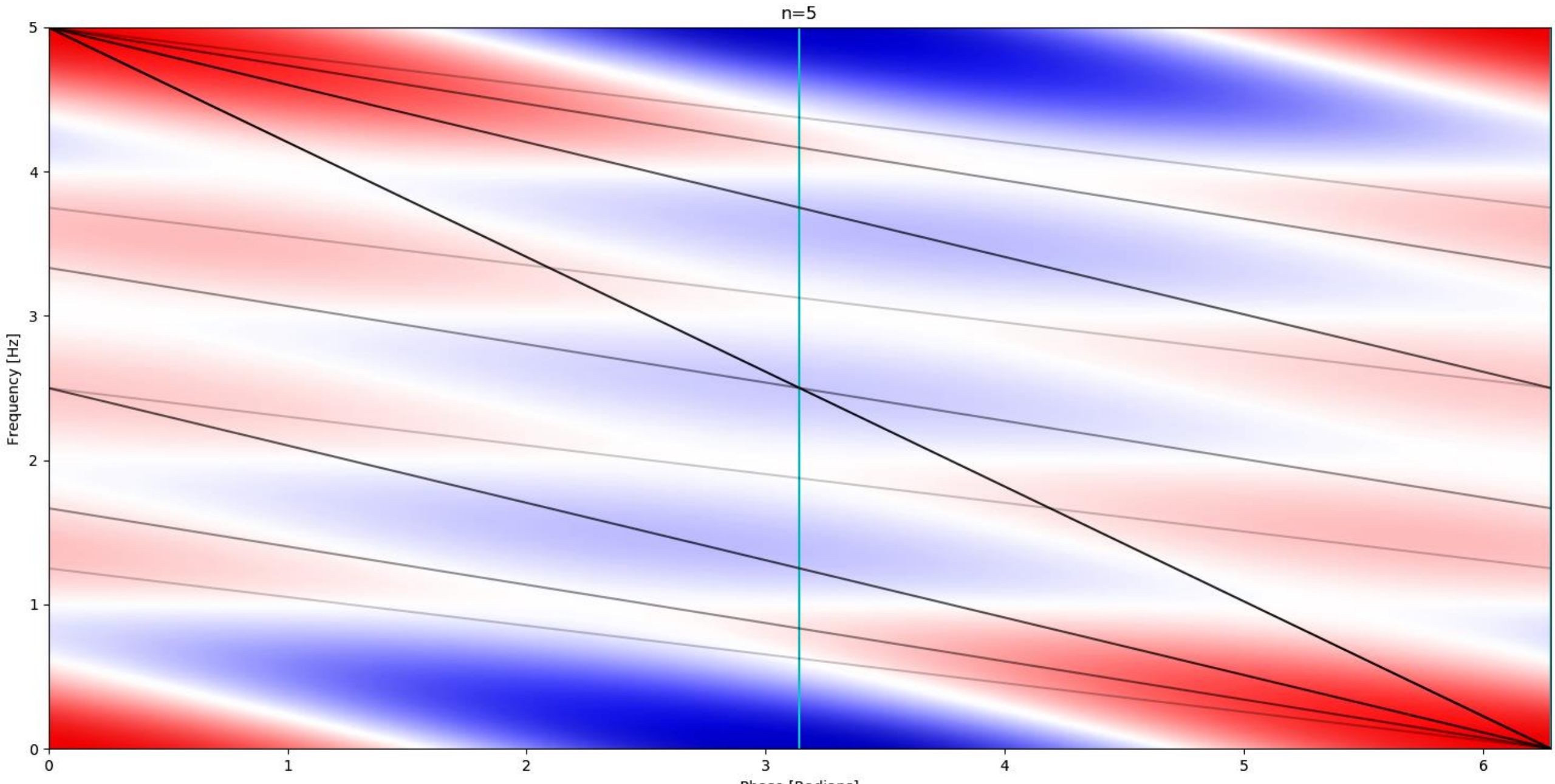
t=4 n=5



PROPOSED NEW TRANSFORM

- Those lines can be interpreted as crests (or valleys, in the case of negative samples) in planar waves describes the deviation of the underlying sinusoid and the sampled value for each point. Samples number 1 to 4 are shown in the image. Below is an example of these planar waves superimposed.

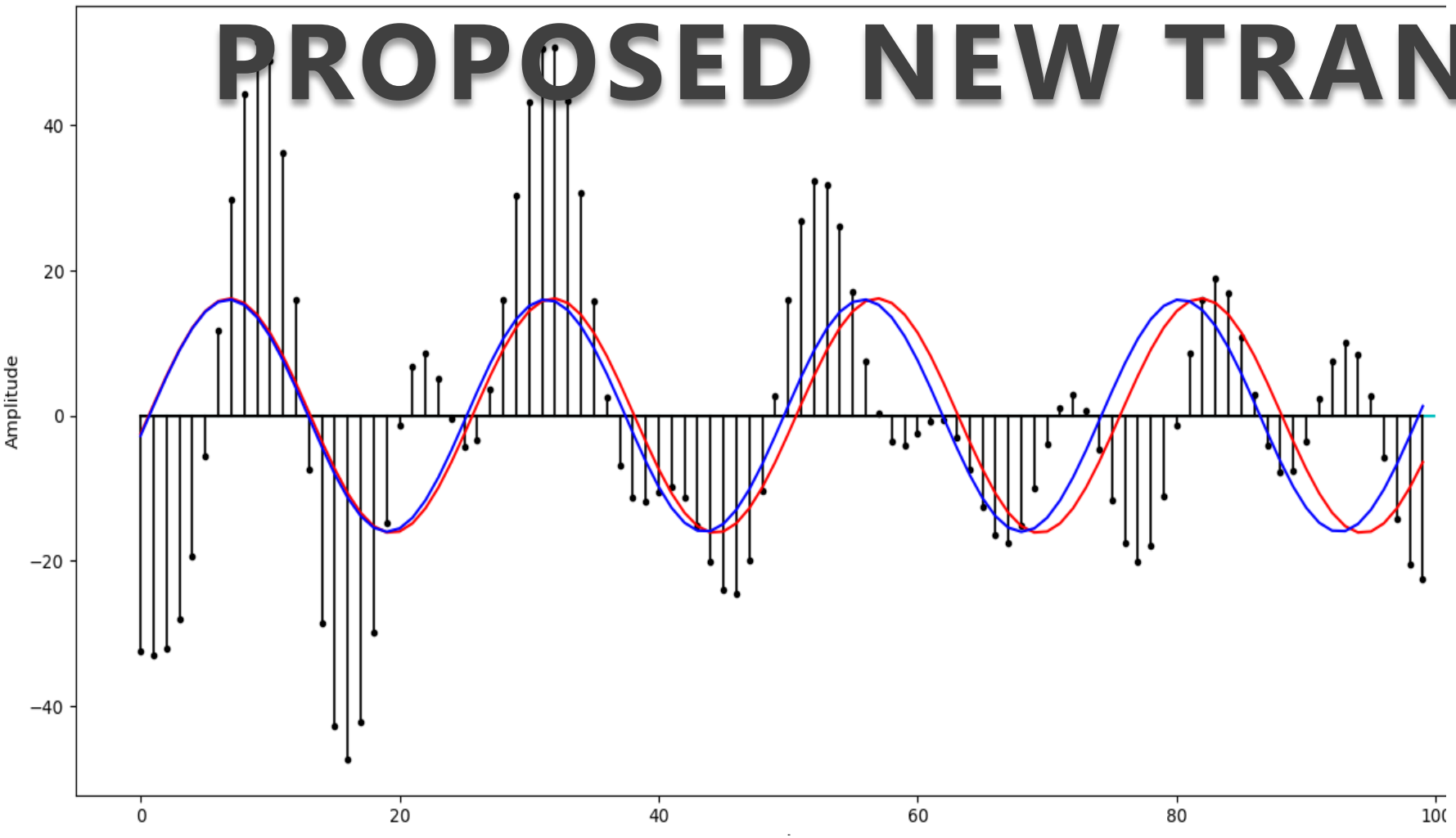
PROPOSED NEW TRANSFORM

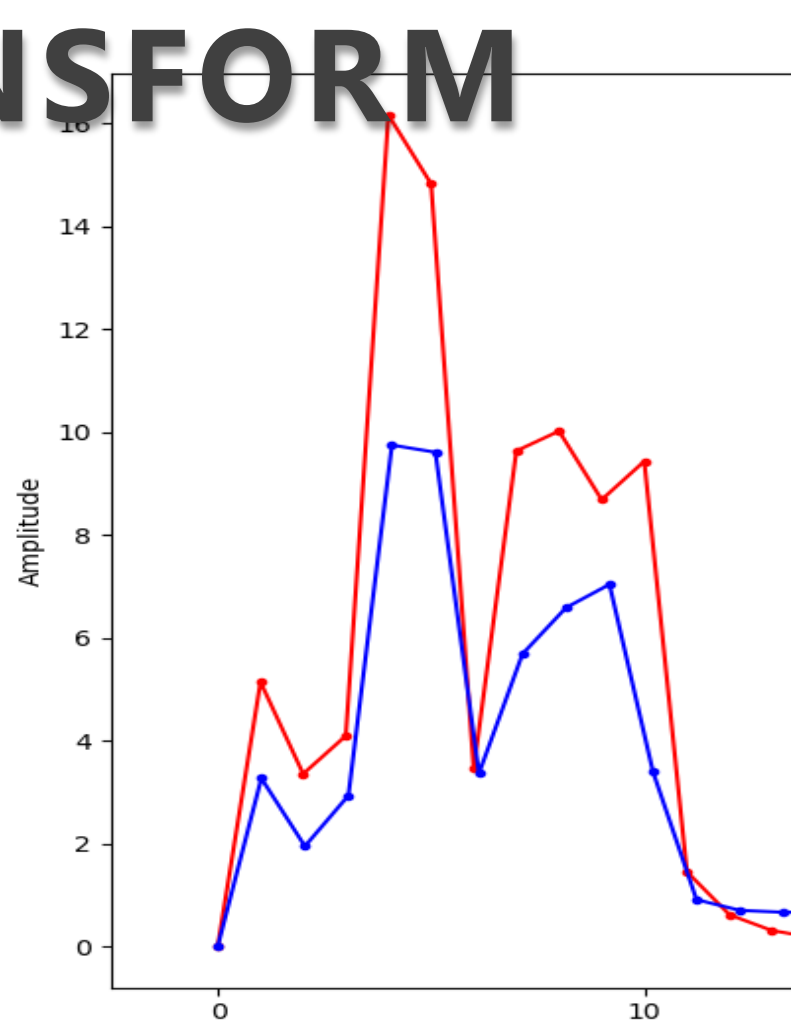


PROPOSED NEW TRANSFORM

- Some useful symmetries can be seen. Besides, if one multiplies each wave by the amplitude of the sample they represent, the superposition of this plot for all waves is identical to the multiplication of a sinusoid with that specific phase and frequency for the signal and, thus, analogous to the Fourier transform results.

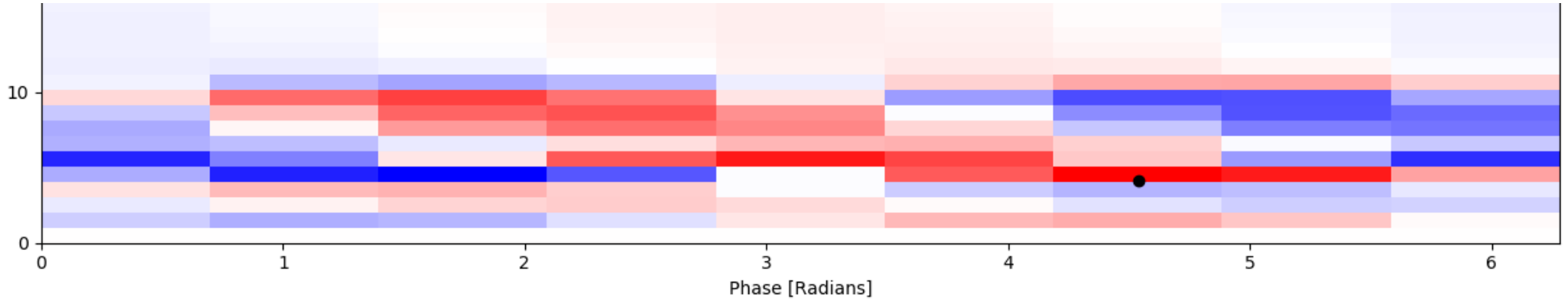
PROPOSED NEW TRANSFORM





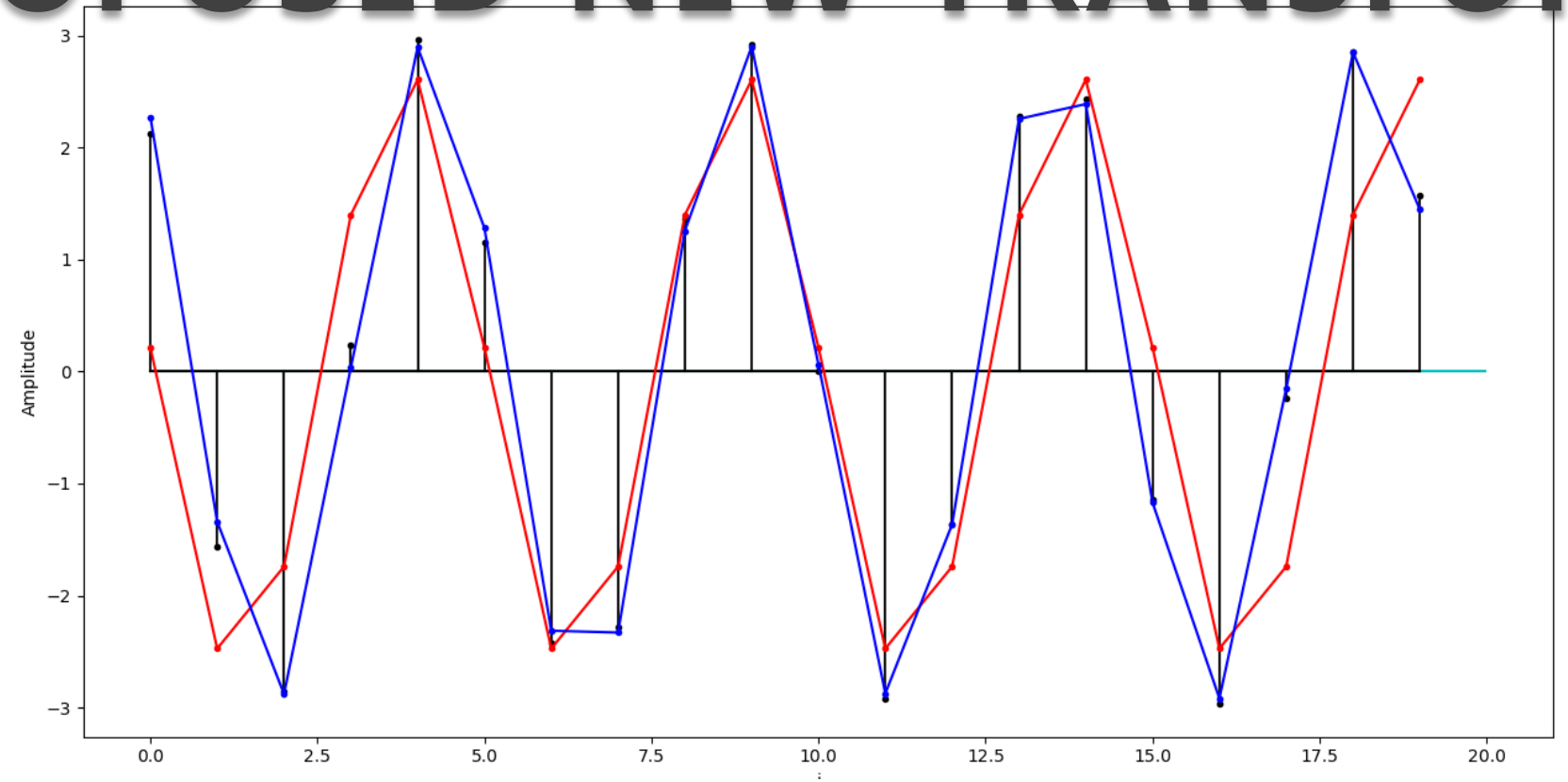
- The image compares this transform with the DFT: in red with have the DFT sinusoid (left) and the underlying discrete signal, and the frequency contents while results in blue are for the proposed transform. The deviation of the sinusoid generated with the novel transform is 1% lower than the DFT one in this particular case.

Proposed new transform



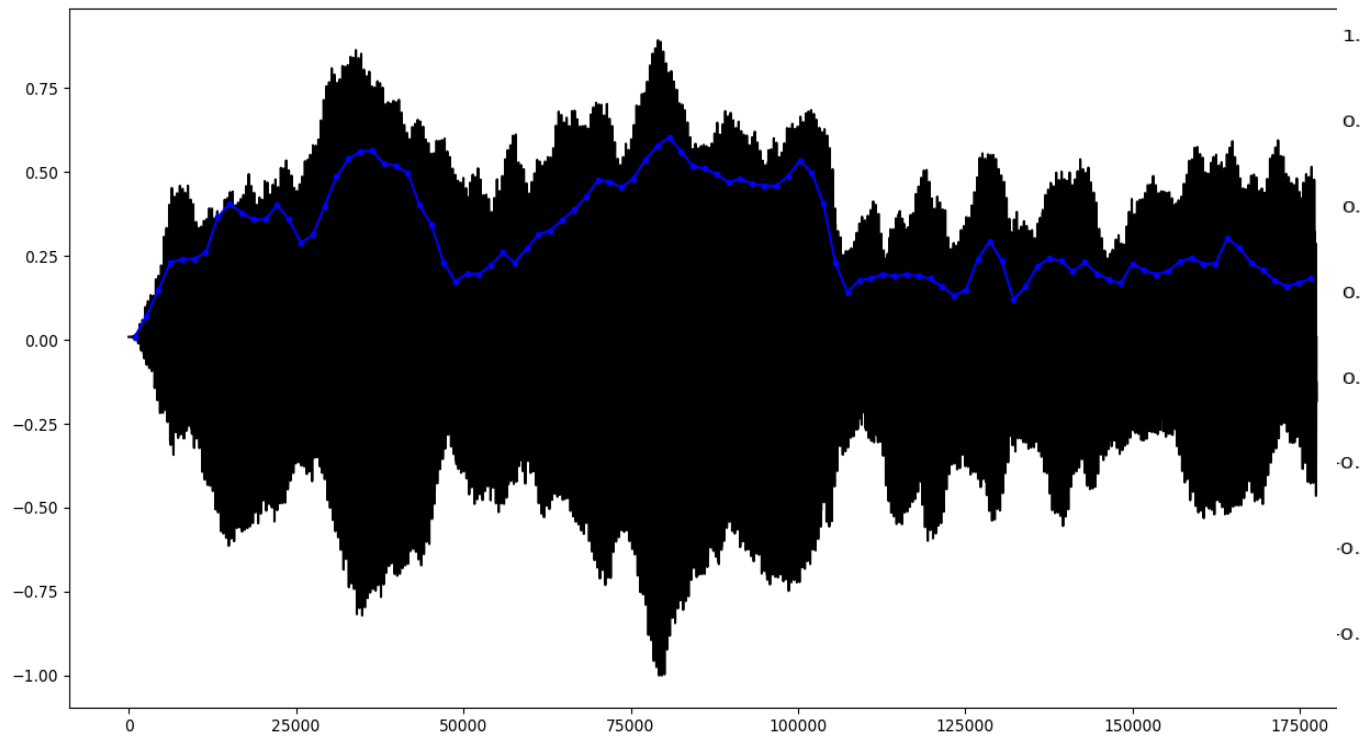
- The result of the transform can be seen above. It's worth noting that, even with a very low resolution, the transform represents an improvement over the DFT, from the point of view of the generated sinusoids. From a computational standpoint, the best FFT implementation have complexity O(n log(n)) while the proposed method can be implement with O(n) algorithms.

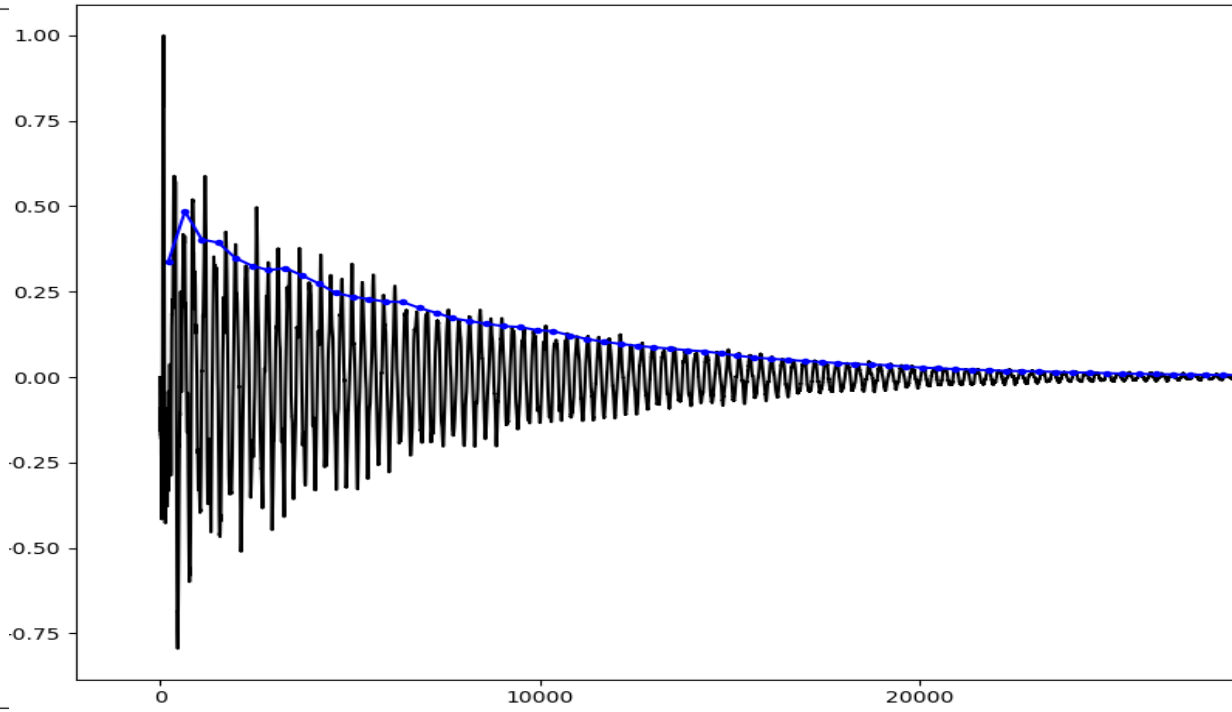
PROPOSED NEW TRANSFORM



- The characteristics of the proposed transform allow a higher accuracy for a small number of samples, as the image illustrates: in red, the DFT approximation has an error 10 times higher than the blue sinusoid, approximated by the proposed transform.

PROPOSED NEW TRANSFORM





- This facilitates the use of the new transform in envelope detection as the above image illustrates for a brass note(left) and the tom of a drum kit (right). Those results were achieved with a preliminary implementation in C++. A usable implementation is expected to be available in the author's github soon. Questions, suggestions, corrections and feedback in general are welcome: tesserato@Hotmail.com

REFERENCES

1 - M-Audio Keystation 49 - https://www.walmart.ca/en/ip/M-Audio-Keystation-49-MK3-USB-MIDI-Keyboard-Controller/PRD4OQTSF97MQQD

2 - ROLI Seaboard Rise 49 - https://www.sweetwater.com/store/detail/SeaboardR-49--roli-seaboard-rise-49

3 - Roland TD-1K - https://www.amazon.co.uk/Roland-TD-1K-Entry-V-Drums-Rubber/dp/B00N5DCJUS

4 - Roland Aerophone GO - https://www.bhphotovideo.com/c/product/1433681-REG/roland_ae_05_aerophone_go_digital_wind.html

5 - FL Studio - https://www.image-line.com/flstudio/

6 - https://commons.wikimedia.org/wiki/File:FFT-Time-Frequency-View.png

7 - WaveNet: A Generative Model for Raw Audio - https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

8 - Deep Voice: Real-time Neural Text-to-Speech - https://arxiv.org/abs/1702.07825

9 - Magenta - https://magenta.tensorflow.org/

10 - Tacotron: Towards End-to-End Speech Synthesis - https://arxiv.org/abs/1703.10135