Characterizing Aural Experiences

Simone Conforti

Simone Conforti conducted his doctoral thesis within the research project Radiophonic Cultures[1] *(2015–2019), which involved the Media Studies department at the University of Basel University, Experimental Radio department of the Weimar University and the Electronic studio at the Music Academy in Basel. He expressed his gratitude to his thesis advisor, the late Prof. Erik Oña, who pushed forward his understanding of music.*

I am going to talk about something that goes in a different direction from what we were discussing in the context of Miro Roman's work. My research is based on classification, or rather, clustering. It is really not about learning, in the sense of machine learning as a generative tool, but trying to understand if there is a way to use machine learning to make the life of musicologists and music researchers easier, in one specific case: sonic classification.

The starting point for the *Radiophonic Cultures* research project was the German experimental radio archive held at the Weimar University. With the sonic research team of the project, I considered using neural networks to create a tool capable of searching the archive without the need for human listening. The first question was, whether we were going to task the neural networks with analysing the sound in whatever form

1 The documentation of the SNSF-funded Sinergia research project Radiophonic Cultures is available at: http://www.radiophonic-cultures.ch/, accessed 21.03.2022

it presents itself, or try a more specific approach. We went for the latter, a specific approach based on the idea of *actual listening*, close to how humans listen to sonic facts. Actual listening means that we preserve an idea of what we heard in a longer time span, giving us the context for a sonic event. We focus on this context in which sonic events are happening, because otherwise sonic events remain unrelated to any musical development. A simple example: let us imagine we hear a voice speaking. We can analyse it in a very short amount of time. This short amount of time can tell us that this is a speaking voice or a singing voice, but it is not telling us anything about the context. Is this a voice speaking alone? Is this voice accompanied by an instrument? Does it belong to a radio drama?

We were searching for a way to represent how humans listen to sonic events, how we relate a sonic source to other sources and understand their meaning. We came up with a very simple idea: instead of analysing content as we would normally do, by slicing sounds in very small chunks and performing a spectrum analysis, we decided to look at bigger frames. A brief overview on the order of scales: if we perform a medium large FFT, we have 4096 bins, which corresponds to a duration of about 93 milliseconds at a sampling rate of 44.1KHz (CD quality). Considering that a 64^{th} note at 60 BPM corresponds to 62.5 milliseconds, we can get an idea about the length of the window from a musical perspective. Humans perceive this short note as a meaningful symbolic representation, but it does not give us enough musical information. If we were to analyse very large FFTs, within which varying sonic events happen, the resulting analysis gives us an average throughout the spectrum. This is interesting with reference to what Selena and Yann told us about Shazam (although the aim is different), the average time of listening chunks in the application being five seconds.[2] Without knowing this, I came to the same conclusion by listening and testing: within five or

2 In the presentation on Radio Explorations by Selena Savić and Yann Patrick Martins as part of this meeting, fingerprints and Shazam are discussed based on the 2003 conference paper by Avery Li-Chun Wang, who developed the Shazam application. The related paper by Wang is available under https://z

six seconds, we have enough time to acquire sufficient information on the kind of music we are hearing, from both high tempo and slow tempo music. For example, if we look at the first six seconds of the 1ere Gymnopédie by Eric Satie, which is in a very slow tempo, this is enough to understand that we are listening to a piano solo source.

In order to learn how to classify these sonic sources, I tried to work with very big FFTs, which requires computational power. In the case of the six seconds the FFT was made of 262144 bins. This was bringing out a lot of information that was complicated to handle, and it was not really representative of the human auditory model. There was not enough reason for making that computational effort.

I started looking for a way to compress this data. I tried many low-level descriptors, and I ended up with the Bark scale representation, which is a way to subdivide the audible frequency range into Critical Bands which are shaped according to the human auditory system. The Bark scale is divided into 24 Bark bands, which can be considered as band pass filters. These band pass filters are tuned on the way the cochlea, our inner ear, subdivides frequencies in space.[3] The cochlea is stimulated by the ear bones, pushing on the oval window – the external part of the cochlea. If we look at the cochlea not as a spiral but unrolled as a sort of cone, we observe that the frequency scale is logarithmically distributed and it is characterised by a more linear tendency towards the lowest frequencies and the opposite towards the high frequencies. This establishes a good relationship between what we hear and a sonic analysis. It is a way to translate the linear scale of the FFT into something that is closer to our auditory model.

When I decided to base my analysis on the Bark scale, I started looking for models that can give me a measure of similarity in order to compare and classify the sonic source. Bark analysis of the large FFTs win-

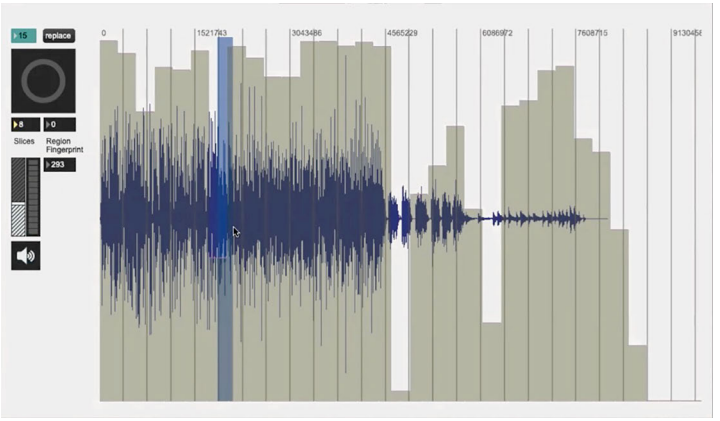
---

enodo.org/records/1416340, and within a Shazam github repository: https://github.com/bmoquist/Shazam

3 Eberhard Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *Journal of the Acoustical Society of America* 33, no. 2 (1961): 248, https://doi.org/10.1121/1.1908630.

dows are performed and compared each to the previous one, using cosine similarity in order to identify where to slice the sections according to their dissimilarities. The Bark scale fingerprints, obtained as the average of the sections, are then used to evaluate the distance from other sources using Euclidian distance. There are other measures of similarity, but this was already giving good results.

Figure 1: The prototype tool for identification of changes in the mix.



Courtesy of Simone Conforti

I have developed a tool prototype, a MAX/MSP patch. It demonstrates a way to measure similarity across audio recordings, and to understand if the idea of using the Bark scale is robust enough. It turned out to be reliable in automatic segmentation, even disregarding the fluctuations in energy of the content. I created a test sound mix. It starts with a radio program, talking with some background music. Then it shifts to dance music, in a more or less similar mood. Afterwards it moves to a solo female singing voice, and then from this moves on to Satie. The loudness is very different in different parts. My tool prototype compares all this through the Bark scale, creating a kind of a step map

that represents the cosine similarity across steps. It identifies points in which we are expecting to have a change in similarity, when the sound content changes. We can then extract a fingerprint and analyse the distance between the chunks. I reduced my Bark scale into 22 bands, probably I can even reduce more, since we are dealing mainly with music or musical sources that present less relevant content beyond 6 or 7KHz. A future step in this research would be to gradually reduce the size of the FFTs, each time we find a change region, to really identify the exact point of change. For the moment, the margin is very large. Still, we can roughly identify the spots in which energy and frequencies characterize a radio talk show, dance music, a singing female voice and the piano. The idea is that we can use these extracted vectors, these fingerprints, as the training set for a self-organizing map. I would use these region fingerprints as the way to organize my map and afterwards I can use whatever input vector as a Bark scale vector to classify a new sound object.

My research so far was attuned to finding ways to classify sound. I think that we can probably reintroduce some more low-level descriptors in order to reinforce this classification. For now, the tool I demonstrated here is quite robust, but sometimes when we have very complex music, it becomes less meaningful or more difficult to interpret. Another relevant topic will be to understand how to shape the results of these comparisons in a way that can be considered meaningful for music researchers.

Finally, I was thinking about possible low-level descriptors for Selena's *Radio Explorations* project. This is also a question to Carl Colena, who understands the particular properties of radio signals and their demodulated audio samples. What do those sounds mean in terms of radio communication? What descriptors could we use to describe those sounds? In the self-organizing map of radio sounds Selena and Yann showed me in May 2020, I saw some inconsistencies, probably because too much was based on the spectrogram. The FFT tells something, but it is not clear what it means. For example, if there is strong energy in the high frequencies, in the case of a recording of instrumental music, perhaps it does not mean so much; it is probably just noise coming from electronic equipment. One question would be whether there is noise in

the radio signal coming from interferences in the transmission, or is it part of the transmission. I was also thinking about the method of using only the FFT representation, with which we can actually get all the descriptions of the sonic source. But how to make good onset detection of an event? It could be performed through the measure of High Frequency Content or through the Spectral Flux, which is another way to not detect the onset but also to see the evolution of a sound within a timeframe. We then have, again, the issue of the time span. Different radio sources present specific patterns, which have different lengths. We can also think of an analysis that describes the source as it happens through Spectral Centroid and Spread, which can be useful because those are characterizing not the frequency in itself, but the energetic perceptual distribution of the sound: the barycentre of the spectrum and the standard deviation. All these features can be considered to define a sound and if needed the analysis can be extended by adding some more audio descriptors in our vectors of features to classify the sound sources.