Ethnographic Sound Recordings Archive Documentation

Release 0.1.2

Michael Blaß, Rolf Bader, Christian Koehn

Apr 02, 2021

CONTENTS

CHAPTER

ONE

ABOUT ESRA

1.1 What is ESRA?

The Ethongraphic Sound Recordings Archive (ESRA) is concerned with the ongoing development of an online archive of digital copies of the extensive collection of rare, often unique, historical ethnographic sound recordings at the Institute for Systematic Musicology (IfSM), University of Hamburg (UHH).

The ESRA is aimed at safeguarding our audio heritage by providing researchers in Systematic Musicology and Ethnomusicolgy as well as cognate disciplines with high-resolution sound files and according metadata in a semantically integrated and easily accessible format.

The collection of historical ethnographic sound recordings at the UHH Institute of Systematic Musicology dates back to the year 1910 and comprises several thousand individual recordings of music and speech from all parts of the world, with a distinct emphasis on the musics of Africa and the Near and Middle East.

It is the stated aim of the ESRA to transfer the sound on these often fragile carriers into the digital domain with the utmost care, developing proprietary methods of playback and digitization to ensure the highest possible degree of fidelity to the source medium. The ESRA furthermore addresses the requirement for sustainable and interoperable access strategies to sound archives' digitized holdings by developing novel methods of computational analysis and advanced formats of representation of the archive's digitized assets to provide researchers with state-of-the-art data to approach the causative questions regarding similarities and divergences hidden in the deep structure of the manifold manifestations of musical expression in the world.

We invite every interested person to sign up for free and have a stab at our services. Any *feedback* is very welcome.

1.2 What is comsar?

Computational Music and Sound Archiving (comsar) is an ongoing project at the IfSM, which develops the computational phonogram archiving standard [[Bad19], [BlassFP20]].

The comsar project also maintains a software package for audio feature extraction. This package provides the main computational backend of ESRA. Please see the software documentation if you are interested in implementational details.

CHAPTER

TWO

FAQ

2.1 Who can register with ESRA

ESRA welcomes every interested person.

2.2 What do I need to register?

To creat an ESRA account you only an valid email.

2.3 What fees do you charge?

ESRA is completely free to use. We do not charge any fees.

2.4 Can I upload my own music?

Yes, you can. See to *upload guide* for further information.

2.5 Which audio formats do you accept?

ESRA supports a variedty of audio formats and codecs. No matter what you choose, it has to be either uncompressed audio or a lossless codec.

2.6 Are there further upload restrictions besides the format?

Yes, audio files must be longer than 30 seconds and shorter than 30 minutes.

2.7 Upload large audio collection

The current version of ESRA supports user uploads for single files. An interface for collection uploads is, however, planned. Nevertheless, *please reach out to us* if you wish to upload a collection.

2.8 Can I make my own song publicly available?

No, publication of user content is currently not supported.

CHAPTER

THREE

TUTORIAL

3.1 Registration

You have to create an ESRA account to explore our audio collections and metadata.

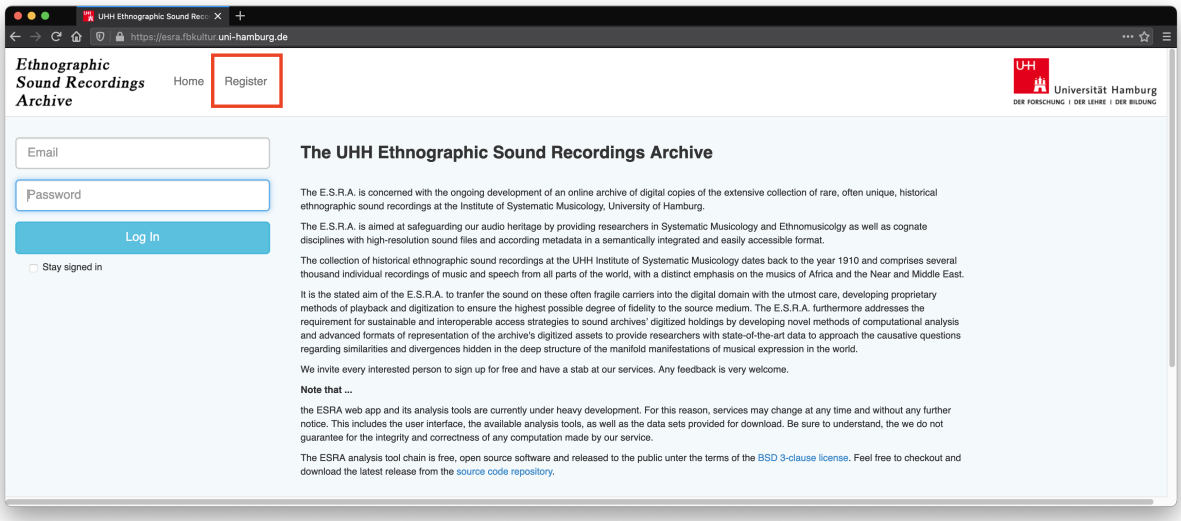


Fig. 1: On the ESRA homepage click on **Register** in the top navigation bar to create an ESRA account.

3.2 Home screen

Once you are logged in you willl see the ESRA home screen. The home screen is the starting point for every interaction with ESRA. It is subdivided into for sections

- menu bar
- meta data search
- public collections
- footer

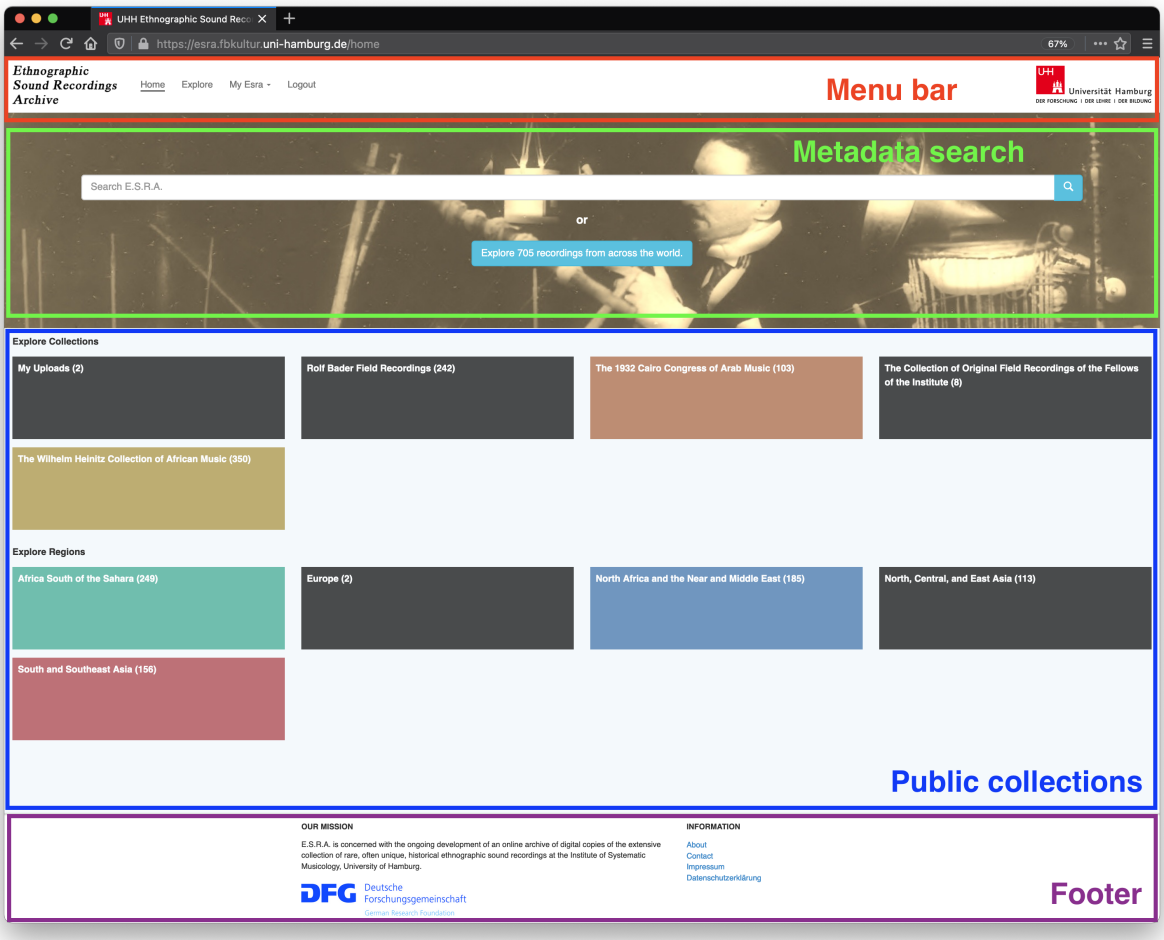


Fig. 2: Sections of the ESRA home screen: menu bar, metadata search, public collections, and footer.

3.2.1 Menu bar

The Menu bar gives you direct access to the main functionallity of ESRA. It remains the same no matter which screen you choose. An underlined navigation link indicates your current position within the application.

- Click on the ESRA logo or the "Home" navigation link to return to the HOME-Screen from anywhere.
- A click on Explore takes you to the Explore-Screen an initializes it with default settings. What these settings are and how you to manipulate them will be explained in the Explore-Screen tutorial.
- Clicking on "My Esra" opens a subnavigation, which takes you to the personal account functionallity, that is, the content upload form, and the private selections. Both menu items will be explained in their corresponding videos.
- The Logout items securely ends your current ESRA session.

3.2.2 Metadata search

Below the menu bar you will find the meta data search bar. This bar enables a common meta data search. Enter a few keywords and hit the "Return" key on your keyboard or press the button with the magnifier glass on it. ESRA will then search for the respective metadata in its database. The results of your query are displayed on the explore screen. ESRA will take you there automatically.

The blue button below the meta data seach bar informs you about the number of records currently registred within the ESRA. Click on it to enter the explore screen with default settings.

3.2.3 Explore by . . .

3.3 Explore screen

The explore screen is devided into two sections: the SOM display on the left and the query result list on the right side.

3.3.1 Query results list

The menus on the top of the song list on the right enable to change between different collections (middle menu), extract subcollections by regions (left menu), or choose a selection (see below).

When searching for an item in the collection with the search text field on the top right, all songs containing the search term in the metadata are displayed. Below is an example of the term 'Kachin' searched for in 'Collection Bader'. The word Kachin does not appear in all song metadata displayed. This is because on this screen, only the most important metadata are shown:

Todo:

- controls
- list click

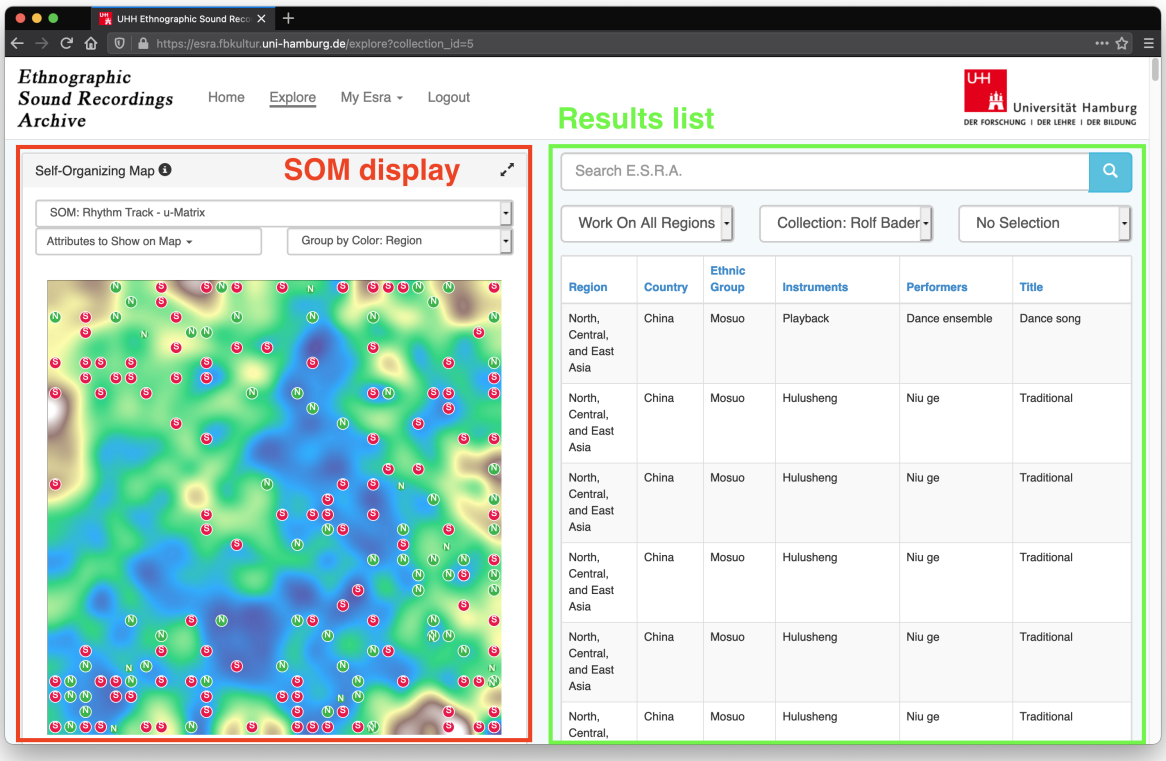


Fig. 3: The right side of the explore screen list search results. Each list item is also diplayed as a circle marker on the SOM disply.

3.3.2 Self-organizing map display

The left side of the explore screen displays different views of the self-organizing map.

3.4 Assett detail screen

3.4.1 Media player

3.4.2 Metadata

3.4.3 Audio features

3.4.4 Referencing

3.4.5 Adding assets to a selection

It is also possible to add an assett to a selection. This is interesting when comparing the results in the SOM. On its assett detail screen, select *Add to my collection*. A list will open, which displays your selection. Click on a selection name to add the current song to it. You can also create a new selection from the same menu.

3.5 Upload music

Important: ESRA supports only uncompressed formats (WAV) and lossless codecs such as FLAC and AIFF. Please make sure that your file meets these requirements before you upload.

To upload content you have to *register with ESRA*. After you are logged in, click on *My Esra* in the navigation bar on the top of the page. In the pop-up menu choose *Upload Private Sound*.

You will be redirected to the upload form. Fill out the form and then click on *Start Upload* at the bottom of the page.

Note: All music you upload, its metadata, as well as the respective audio features as extracted by ESRA are private to your account. Other users cannot see or play your music, or access it in any other way.

Audio feature extraction will start immediately after the upload has completed. Depending on the size of your file, extraction routines may take several minutes. You can check the status of the feature extraction on the *assett detail screen*.

3.6 Create a selection

3.7 Export audio features

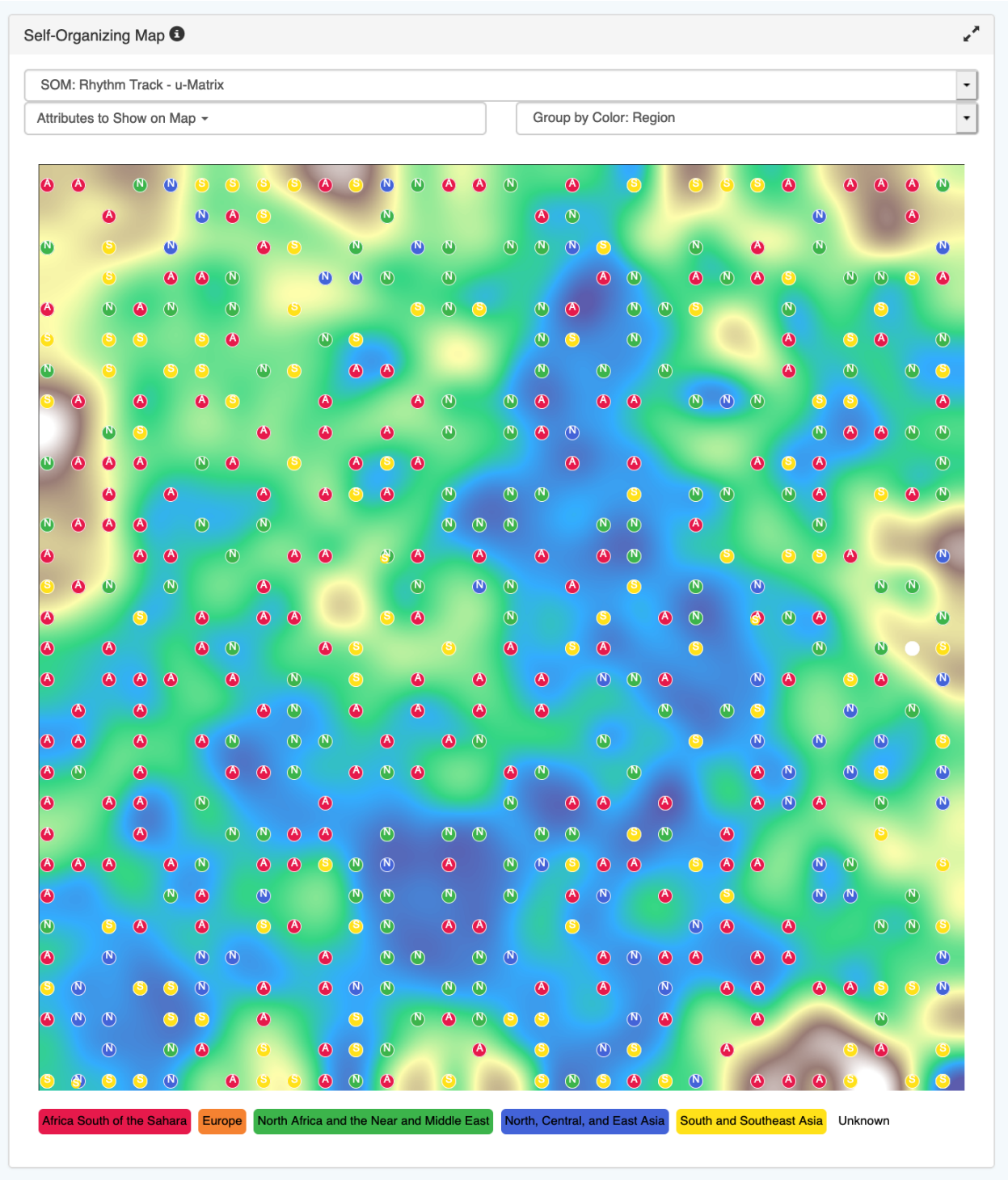


Fig. 4: The SOM display in ESRA including visualization controls.

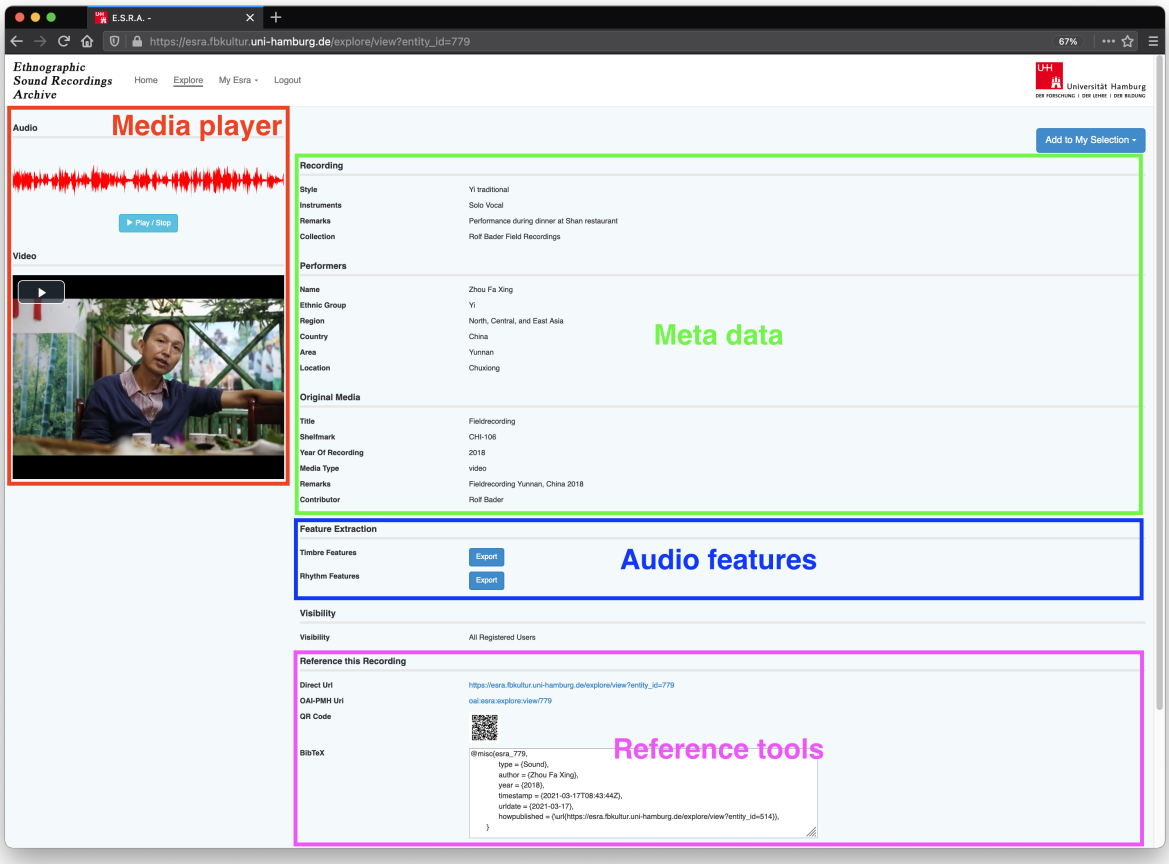


Fig. 5: The assett detail screen includes a media player, all available metadata, access to audio features, and tools for referencing the selected recording.

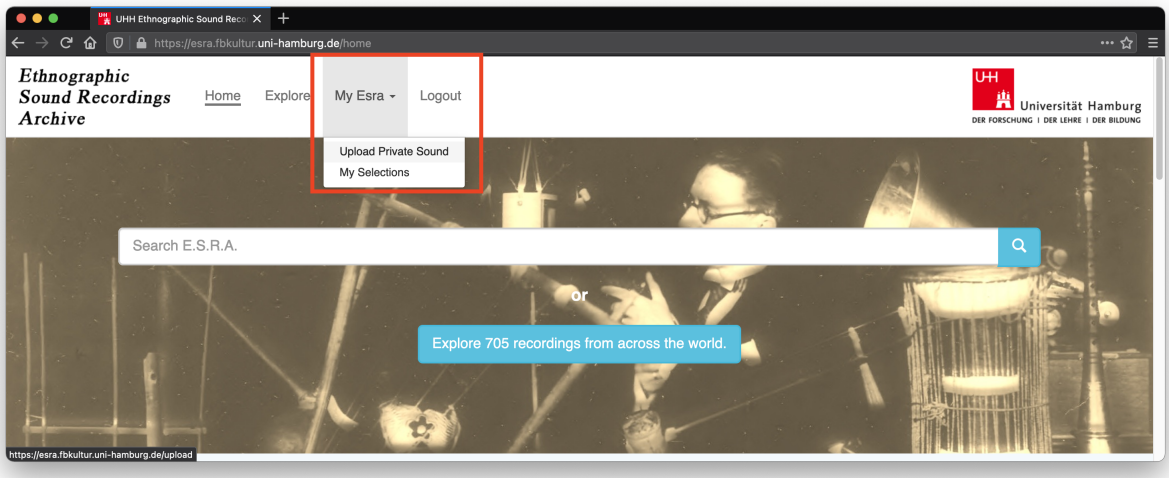


Fig. 6: Click on **My Esra** then on **Upload Private Sound** to navigate to the upload form.

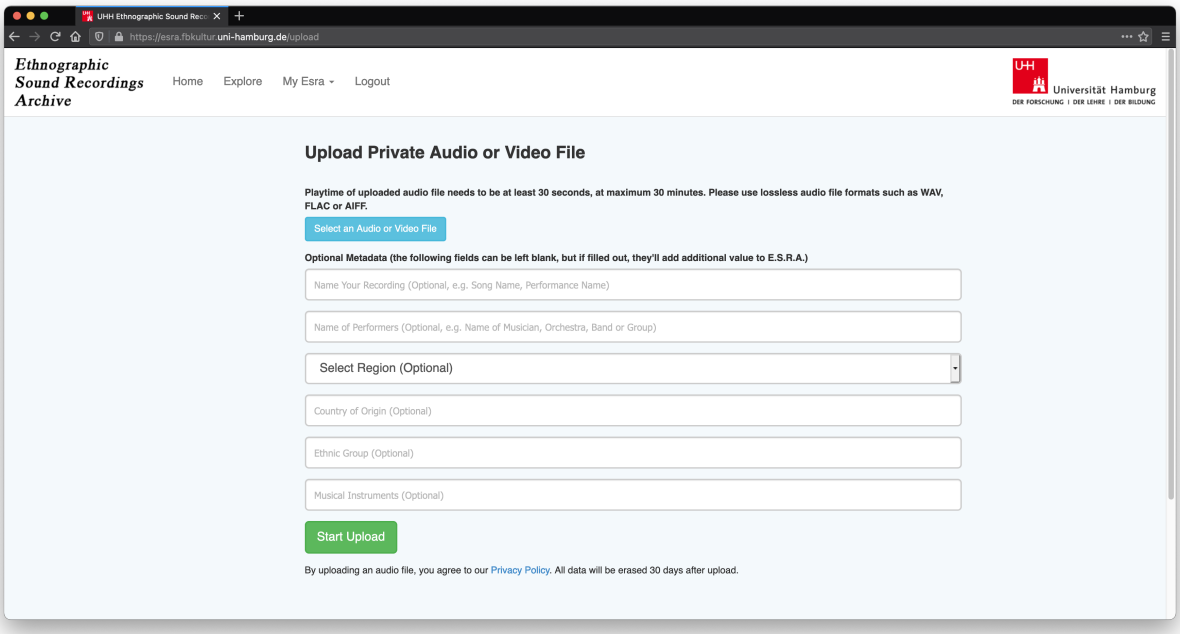


Fig. 7: Fill out the upload form to upload sounds into ESRA. Metadata are optional.

CHAPTER

FOUR

TRACK SYSTEM

4.1 Pitch Track

The pitch track of the COMSAR framework extracts pitch from a soundfile, accumulates pitches into an octave, detects notes, vibrato, slurs, or melisma, determines most likely tonal systems, extracts the melody, and calculates n-gram histograms.

The pitch track instance should contain the frame size and hop ratio. In the example notebook 100 pitches per second are analyzed (see jupyter example notebook **COMSAR_Melody_Example.jpynb**)

4.1.1 Melody

An example of a Lisu solo flute piece (ESRA, vimeo) for pitch and melody extraction is shown in the figure below.

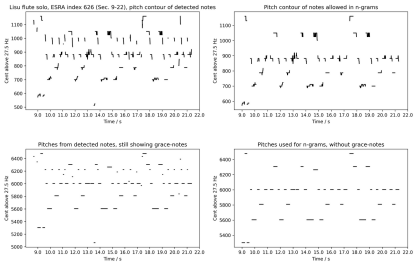


Fig. 1: Example of pitch and melody extraction using Lisu flute solo, ESRA index 626. The analysis has five stages of abstraction. 1) pitches are calculated over the whole piece. From pitch values to melodies: 2) top left: Pitch contours of detected notes, 3) top right) pitch contour of notes allowed for n-grams (melodies), 4) bottom left: mean pitches of allowed notes from plot 2) still showing grace-notes, 5) bottom right: mean pitches of notes allowed for n-grams.

In a first analysis stage, the original sound files are analyzed with respect to pitch, where n pitch values are calculated per second, determined by the args used in the instantiation of the pitch track. Pitch analysis is performed using the autocorrelation method, where the first peak of the autocorrelation function determines the periodicity of the pitch.

For frequencies below about 50 Hz the result of the autocorrelation is not precise. In nearly all cases, the amount of period cycles the autocorrelation integrates over is not an integer, but a fraction of one period is integrated over at the end of the sound. With higher frequencies, this is not considerable. Still, with low frequencies, due to the few sound period cycles the autocorrelation integrates over, this leads to incorrect results. Here an algorithm is used to compensate for overintegration. For frequencies above 1.7 kHz, results again are not correct due to the limited sample frequency, allowing only certain periods. Here, oversampling is performed to compensate for incorrectness. Still, in most cases, melody falls into the human singing range, and, therefore, the algorithm does not need these corrections.

The pitches are transferred into cent, using a fundamental frequency of f0 to be determined in the args (f0 = 27.5 Hz is recommended for sub contra A) in noctaves above f0 (noctaves = 8 is recommended), and a cent precision dcent (dcent = 1 cent is recommended).

The second abstraction stage uses an agent-based approach, where **musical events, notes, grace-notes, slurs, melismas, etc.** are detected. The agent follows the cent values from the start of each musical piece and concatenates adjacent cent values according to two constraints, a minimum length (minlen) a pitch event needs to have and a maximum allowed cent deviation (mindev). So, e.g., with pitch track instantiated using 100 pitch values per second, a minlen = 3 is 30 ms minimum note length, and mindev = 60 his allows for including vibrato and pitch glides within about one semitone, often found in vocal and some instrumental music. Lowering the allowed deviation often leads to the exclusion of pitches, which often have quite strong deviations. As an example, an excerpt of 23 seconds of a Lisu solo flute piece is shown in the figure below on the top left. Some pitches show a quite regular periodicity, some are slurs or grace-notes.

The third abstraction stage determines **single pitches** for each detected event by taking the strongest value of a pitch histogram. As can be seen in the top left figure, often pitches are stable, only to end in some slur in the end. Therefore, taking the mean of these pitches would not represent the main pitch. Using the maximum of a histogram, on the other side, detects the pitch most frequency occurring during the event. On the bottom left, this is performed and can be compared to the top left plot. When listening to the piece, this representation seem to contain still too many pitch events. So,. e.g., the events around 6000 cent are clearly perceived as notes. Still, those small events preceding around 6200 cent are heard as grace-notes. Therefore, to obtain a melody without grace-notes, a fourth stage needs to be performed.

In a fourth stage, pitch events are selected using three constraints to allow for **n-gram construction**. n-grams have shown to represent melodies stable and robust in terms of melody identification, like e.g., in query-by-humming tasks. Here, n adjacent notes are clustered in an n-gram. A musical piece, therefore, has N-n n-grams, where N is the amount of notes in the piece. The n-grams are not constructed from the notes themselves, but from the intervals between the notes. Therefore, a 3-gram has two intervals. Also, the n-grams are sorted in 12-tone just intonation. Therefore, each interval is sorted into its nearest pitch-class. Further implementations might include using tonal systems as pitch classes. Usuall,y 2-grams or 3-grams are used, sometimes up to 5-grams. All n-grams present in a piece are collected in a histogram, where, in the present study, the nngram most frequent n-grams (ten in this case) are collected into a feature vector to be fed into the machine learning algorithm.

So adjacent notes qualifying for n-gram inclusion need to be such to exclude grace-notes, slurs, etc. This is obtained by demanding the notes to have a certain length (minnotelength in amount of analysis frames), in the example below 100 ms is used (minnotelength = 10 as pitch track instantiation with 100 pitches per second was performed). Additionally, a lower (ngcentmin) and upper (ngcentmax) limit for adjacent note intervals is applied. The lower limit is 0 cent in the example to allow for tone repetition. The upper limit is set to +-1200 cent here, so two octaves, most often enough for traditional music. This does not mean that traditional pieces do not have larger intervals, as e.g., expected in jodeling. Still, such techniques are not used in the present music corpus, and even when present, they are not expected to be more frequent than smaller intervals. In the top right figure, we see the pitch contours for all allowed notes used in n-gram calculation. Indeed, all grace-notes are gone.

In a last step, shown in the bottom right plot, the pitches of each event are again taken as the maximum of the histogram of each event. Now following the musical piece aurally, the events represent the **melody**. These notes are used for the n-gram vector.

An example of a trained SOM using musical pieces of Kachin ethnic group, northern Myanmar with Uyghur musical pieces from Xinjiang, western China is shown below. The map is trained with 5-grams. Most Uyghur pieces are located at the lower blue region, mainly because of the frequent tone repetitions of Uyghur music compared to Kachin songs.

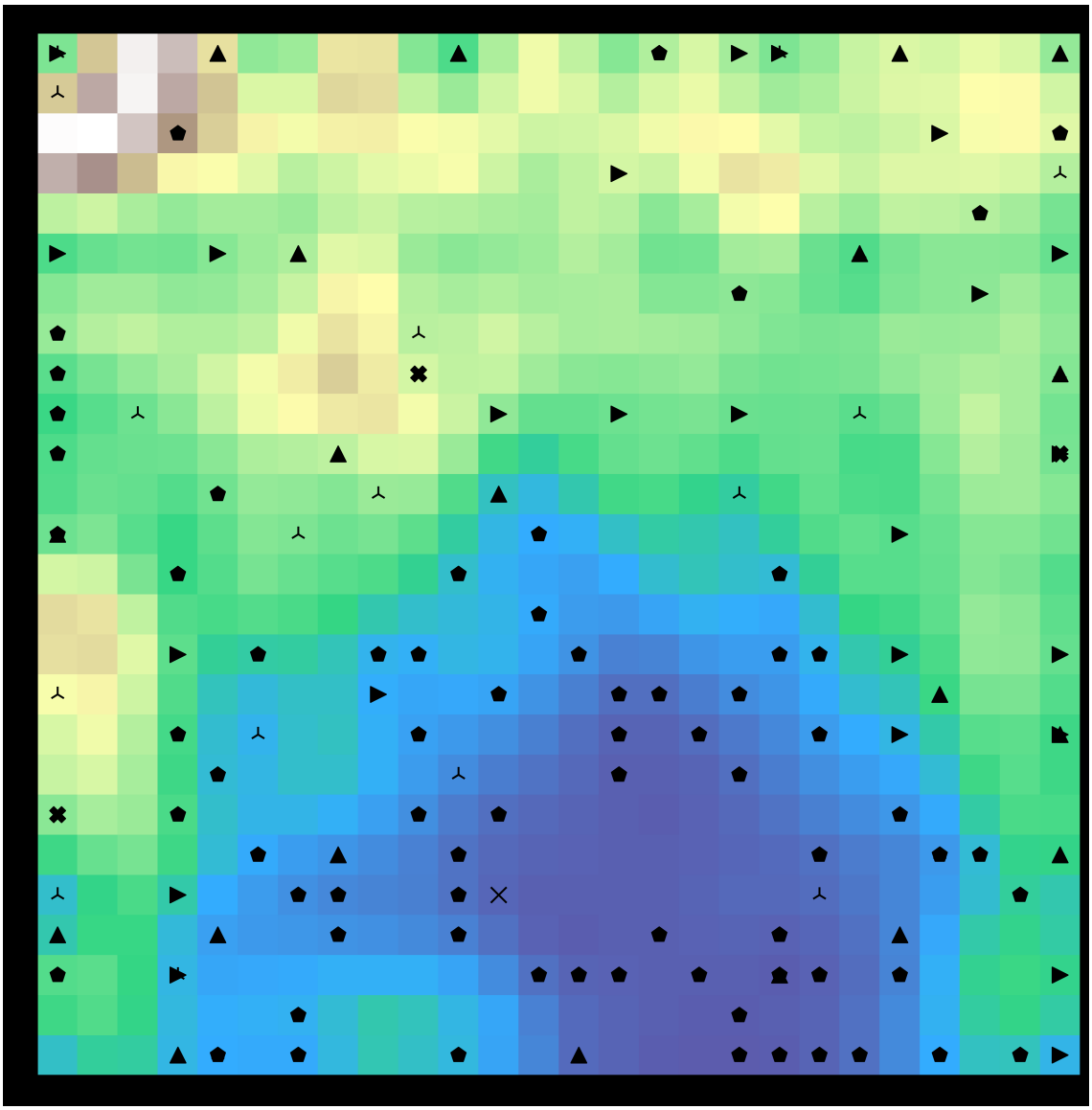


Fig. 2: Trained 5-gram SOM, comparing Kachin and Uyghur musical pieces. Most Uyghur pieces are located in the lower blue region due to enhanced note repetition found in Uyghur music.

4.1.2 Tonal System

Tonal systems are normally understood as a small set of cent values. So a 7-tone scale has seven-pitch values. Still, depending on musical performance, determining a tonal system is more or less complex. Articulation in singing often leads to a large variation in pitch. The same might hold for lutes, guitars, violins, or wind instruments. Percussion instruments have a much more straight pitch. till, they often are inharmonic, and, therefore, a pitch might not even be perceived.

Therefore the MIR tool for investigating tonal systems takes tonal systems as an accumulation of pitch values over mainly single-voiced musical pieces compressed within one octave with a precision of one cent. An autocorrelation algorithm determines the pitch for n time frames per second, the one already shown in the melody section above.

In a second stage, pitch events are detected, again as discussed above. All pitches of the detected musical events are then accumulated in dcent values (dcent = 1 is recommended), starting again from f0 in noctaves. To also include melismas and slurs in the calculation, the tonal system is derived from all pitch values in the musical event (top left plot of above figure). If the tonal system should only be constructed from pitch events with a very constant pitch, the mindev parameter needs to be small.

The strongest frequency maxf is then taken as the fundamental of the tonal system. In the tonal system plots shown below, this lowest cent is not shown, as it often overwhelms the other accumulated cent values.

In a last step, using the largest accumulated pitch as fundamental of the tonal system, all accumulated cents in noctaves are mirrored into one octave. When a precision of one cent was used, the input feature vector to the SOM has a length of 1200, reflecting 1200 cent in one octave.

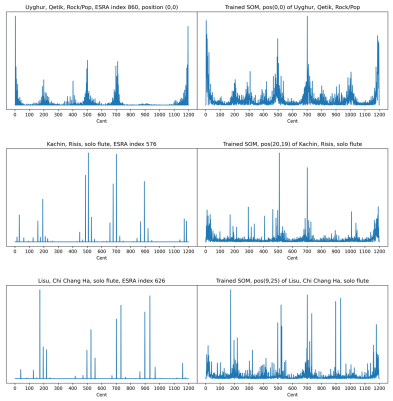


Fig. 3: Three examples of tonal systems as calculated from a sound file (left column) and as a vector on the neural map on the location the musical pieces fits best (right column). Top: Uygur Rock/Pop piece by Qetik, Middel: Kachin flute solo piece, Bottom: Lisu flute solo piece.

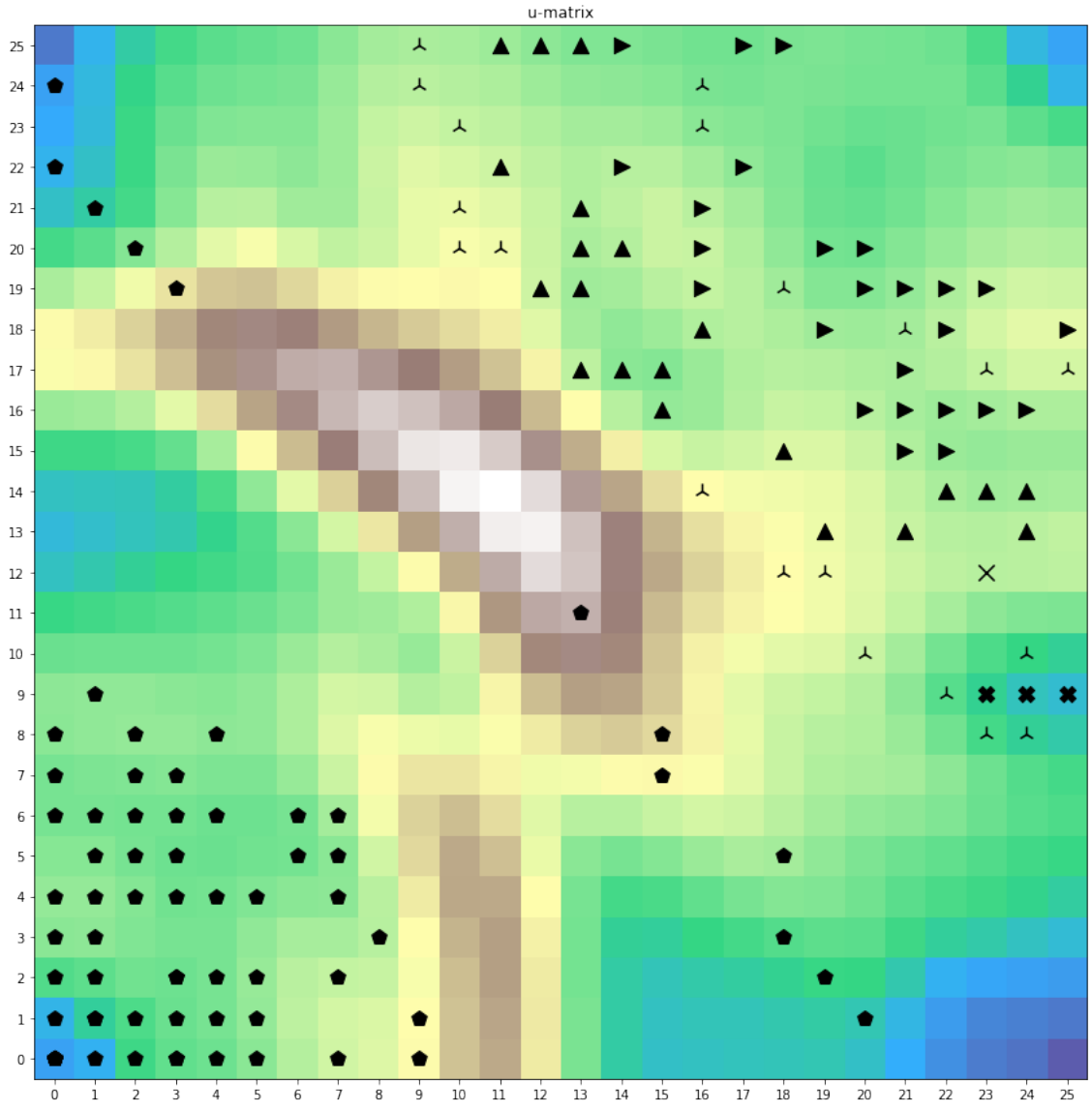
The outputs are

- accumulated cent values over noctaves
- accumulated cent values within one octave
- Names of the ten best-matchin tonal systems
- Cent values of the best-matching tonal systems
- Correlation of each cent value in all best-matching scale to estimate the salience of each note to the overall large correlation between theoretical scale and calculated values
- Overall correlation of the best-matching scales

The tonal systems used for comparison are taken from a set of scales: https://www.flutopedia.com/scales.htm. The list contains over 900 scales. Therefore, matches might not meet expectations. Reduced lists fitting special intrests will

be developed in the future.

Below, a trained SOM with tonal systems of the Kachin vs. Uyghur music case is shown below. Both ethnic groups are clustered, where Uyghur pieces on the lower left are nearer to just intonation, while Kachin shows more deviating pitches.



4.2 Rhythm Track

ESRA includes musical pieces from more than 70 ehtnic groups collected in over 50 countries. A reasonable system for rhythm similairity estimation has to

The rhythm track implemets a timbre-based theory of musical rhythm Blaß [[Blass13]] [Blaß2019]). That means, it does not address the relative temporal distance between note onsets. Instead it focuses on the actual sounds that are played. To this end, ESRA runs an *onset detection* algorithm, which estimates the temporal position of note onsets within the piece under consideration. Thereafter, *audio features extraction* computes a measure for the perceived

brightness of the sounds at each onset. ESRA then estimates the probabilites to change from one given sound to another. This analysis is carried out using a *Hidden Markov Model*.

This approch to rhythm has several advantages:

- Since the model operates only on the actual sound, it avoids any cultural bias. Especially, the model does not apply notions of Western rhythm theories to other music cultures.
- The numerical representatin of the rhythm is of exactly the same size for each piece analyzed in the same track no matter how long the actual piece is. This is a crucial feature for further processing stages.
- Musical from literally any music culture a be compoared on a well-defined basis.

However, the model does have a some disadvantages, too. These include first, that temporal information is reduced to a equidistant succession. Additionally, the model as such can be "hard to read" for humans. This fact is, however, mitigated by system structure. Since the models are automatically compared, users only have to interprete the ouput of the similarity estimation, which is straight foreward. They, hence, do not have to care too much about details of the rhythm model.

4.2.1 Onset detection

Onset detection estimates the starting points of note onsets in digital audio signals. Our approach is based on the standard spectral flux method. An onset is assumed between two consecutive STFT segements if there is a high per band difference in spectral energy.

Further details can be found in the documentation of the spectral_flux() method and the FluxOnsetDetector class from the apollon framework.

4.2.2 Audio feature extraction

ESRA descrbies the timbre of an onset in terms of the perceived brightness. The spectral centroid is know to correlated well with the perception of brightness.

4.2.3 Hidden Markov Model

Work in progress.

4.3 Timbre Track

The TimbreTrack combines a set of audio features to model the timbral content of piece of music. These features are computed on short, consecutive portions of the input signal.

4.3.1 Frequency domain features

Frequency domain features are computed from a spectrogram. A spectrogram displays the distribution of energy within the audible frequency bands of consecutive, short portions of the input signal. ESRA computes its spectrograms using the Discrete Fourier Transform.

The shape of each spectral distributions is related to how humans perceive the timbre of the related portion of a musical piece. Typically, the first four moments of a distribution are utilized to describe its shape.

Spectral Centroid

The first moment of a spectral energy distribution is the spectral centroid frequency. It is a measure of central tendency. It marks the frequency that is considered as the center of a spectrum. It may be computed as the weighted arithetic mean of a frequency spectrum.

Several studies could confirm that the spectral centroid correlates strongly with the human auditory perception of *brightness*. Moreover, brightness is the most salient dimension of timbre perception.

Detailed information can be found in apollon.signal.features.spectral_centroid.

Spectral Spread

Spectral Spread is the second moment of a spectral distribution and refers to its variance. It is a measure of how much frequencies deviate from the spectral centroid frequency.

Unlike Spectral Centroid, Spectral Spread does not map immediately to a perceptional quality.

Detailed information can be found in apollon.signal.features.spectral_spread.

Spectral Skewness

Spectral skewness is the third spectral moment. A skewness of 0 means that the distribution is perfectly symmetric. Negative values indicate a bias towards high frequency. Conversely, positive values indicate a displacement to low frequencies.

Detailed information can be found in apollon.signal.features.spectral_skewness.

Spectral Kurtosis

Spectral kurtosis is the fourth spectral moment. It is a measure for the shape of the tails of a distribution.

Detailed information can be found in apollon.signal.features.spectral_kurtosis.

Spectral Flux

Spectral flux is the change in amplitude per frequency bin over time. It is particularly useful for timbre.

Detailed information can be found in apollon.signal.features.spectral_flux.

4.3.2 Time domain features

Fractal correlation dimension

The fractal correlation dimensions measures chaoticity. Chaoticity is defined the the amount of harmonic overtone spectra plus large amplitude fluctuations. So a single guitar tone in its steady-state has a fractal correlation dimension of one. A piano chord with three keys pressed has a fractal dimension of three. Each inharmonic sinusoidal added another dimension. White noise therefore has a dimension of infinity.

To calculate a fractal correlation dimension, a time series x(t) of recorded sound with n sample points is embedded in a pseudo phase-plot like

$$X(t) = \{x(t), x(t+\delta), x(t+2\delta), ..., x(t+d\delta)\} .$$

Starting from X(t) we then calculate the 'area' or 'volume' C(r) like

$$C(r) = \frac{1}{N^2} \sum_{i \neq j} H(r - |\mathbf{X}_i - \mathbf{X}_j|) \, .$$

Here, H(x) is the Heavyside function with

$$H(x) = \begin{cases} 0, & \text{for x} \leq 0 \\ 1, & \text{for x} > 0 \end{cases}$$

which counts the amount of points within the radius r. C(r) is a probability, as we normalize with respect to all N^2 possible combinations.

Then the correlation dimension is defined as

$$D_C = \lim_{r \to 0} \frac{\ln C(r)}{\ln r} \, ,$$

which is derived from the idea of the definition of the dimension (Bader2013). The exponent is the dimension which is the slope of a log-log plot. So practically we need to take the middle of the distribution and look for a constant slope in the log-log plot. This slope is the correlation dimension.

This is a very powerful tool as it has certain properties:

1. If only one harmonic overtone spectrum is in the sound, DC = 1 no matter how many overtones are present.
2. Each additional harmonic overtone spectrum raises DC to the next integer.
3. If only one inharmonic sinusodial is added, DC raises to the next integer making it suitable for detection of additional single inharmonic components too.
4. Large amplitude fluctuations lead to a raise of DC.
5. As the absolute amplitude is normalized, DC does not depend upon amplitude.
6. If a component is below a certain amplitude threshold it is no longer considered by the algorithm.

The fractal correlation dimension raises with initial transients, as they contain chaoticity. It is also a good measure of event density in a musical piece.

4.3.3 Models of perceptional qualities

Loudness

Although several algorithms of sound loudness have been proposed by Fastl and Zwicker [[FZ07]], for music still no satisfying results have been obtained [[vR13]]. Most loudness algorithms aim for industrial noise and it appears that musical content considerably contributes to perceived loudness. Also loudness is found to statistically significantly differ between male and female subjects due to the different constructions of the outer ears between the sexes. Therefore a very simple estimation of loudness is used, and further investigations in the subject are needed. The algorithm used is

$$L = 20 \log_{10} \frac{1}{N} \sqrt{\sum_{i=0}^{N} \frac{A_i^2}{A_{ref}^2}} \, .$$

This corresponds to the definition of decibel, using a rough logarithm-of-ten compression according to perception, and a multiplication with 20 to arrive at 120 dB for a sound pressure level of about 1 Pa. Of course the digital audio data are not physical sound pressure levels (SPL), still the algorithm is used to obtain dB-values most readers are used to. As all psychoacoustic parameters are normalized before inputting them into the SOM, the absolute value is not relevant.

Roughness

Roughness calculations have been suggested in several ways (for a review see [[SvRB09]] and [[Bad13]]). Basically two algorithms exist, calculating beating of two sinusoidals close to each other [[SvRB09], [Set93], [vH63]] or integrating energy in critical bands on the cochlear [[FZ07], [Sot94]]. The former has been found to work very well with musical sounds, the latter with industrial noise.

In the paper a modified Helmholt/Bader algorithm is used. Like Helmholtz it assumes a maximum roughness of two sinusoidals at 33 Hz frequency difference. As Helmholtz did not give a mathematical formula how he did calculate roughness, according to his verbal descriptions a curve of the amount of roughness R_n is assumed between two frequencies with distance df_n which have amplitudes A_1 and A_2 like

$$R_n = A_1 A_2 \frac{|df_n|}{f_r e^{-1}} e^{-|df_n|/f_r} .$$

with a maximum roughness at $f_r = 33$ Hz. The roughness R is then calculated as the sum of all possible sinusoidal combinatins like

$$R = \sum_{i=1}^{N} R_i .$$

The goal of Schneider *et al.* [[SvRB09]] was to model the perceptual differences of tuning systems like Pure Tone, Werkmeister, Kirnberger, etc. in a Baroque piece of J. S. Bach. This task required high temporal and spatial resolution in the frequecy domain. The authors, therefore, utilized a Discrete Wavelet Transform (DWT). The roughness analysis in ESRA does not aim for such subtle differences, but for an overall estimation. Moreover, ESRA has to accomodate resource restrictions. To this end, the DWT was replaced by the Discrete Fourier Transform.

Sharpness

Perceptual sharpness is related to the work of Bismarck (Bismarck1974) and followers (Aures1985b, Fastl2007). It corresponds to small frequency-band energy. According to (Fastl2007) it is measured in acum, where 1 acum is a small-band noise within one critical band around 1 kHz at 60 dB loudness level.

Sharpness increases with frequency in a nonlinear way. If a small-band noise increases its center frequency from about 200 Hz to 3 kHz, sharpness increases slightly, but above 3 kHz strongly, according to perception that very high small-band sounds have strong sharpness. Still sharpness is mostly independent of overall loudness, spectral centroid, or roughness, and therefore qualifies as a parameter on its own.

To calculate sharpness the spectrum A is integrated with respect to 24 critical or Bark bands, as we are considering small-band noise. With loudness L_B at each Bark band B, sharpness is

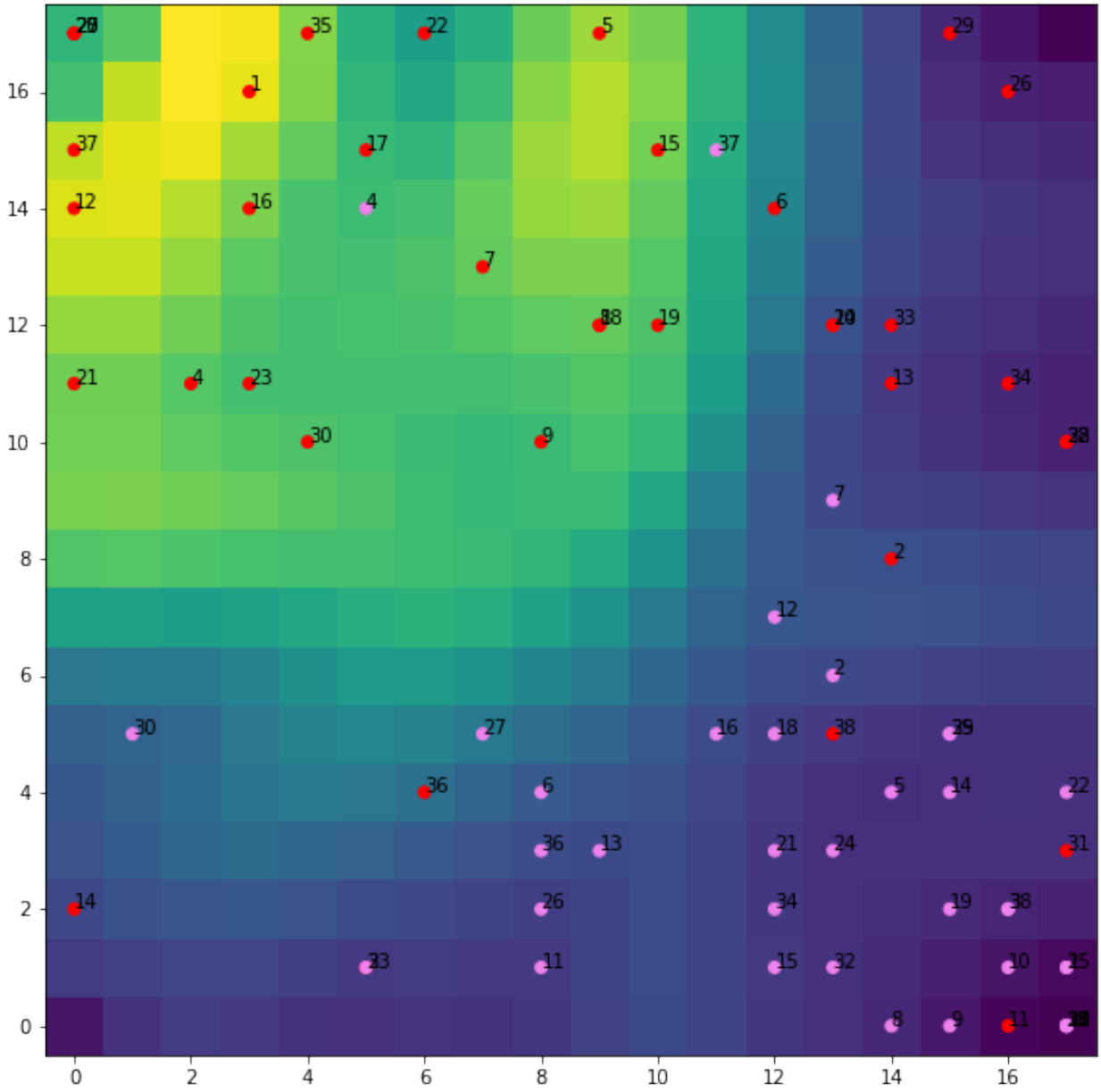
$$S = 0.11 \frac{\sum_{B=0}^{24Bark} L_B g_B B}{\sum_{B=0}^{24Bark} L_B} ,$$

where a weighting function g_B is used strengthening sharpness above 3 kHz like({Peeters2004}

$$g_B = \begin{cases} 1 \text{ if} B < 15 \\ 0.066 e^{0.171B} \text{ if} z \geq 15 \end{cases}$$

4.3.4 Example

Below the an example result of two sets of Hip Hop musical pieces. A SOM was trained with a certain amount of features, able to cluster Chinese (red) and Western (violet) Hip Hop pieces.



CHAPTER

FIVE

SELF-ORGANIZING MAPS EXPLAINED

The self-organizing map (SOM) is like a brain that has learned the musical pieces of a collection. The ESRA SOM in ESRA is trained on all pieces in ESRA. You can train your own SOM in the offline version.

Online, there are two trained SOMs, one for timbre and one for rhythm (for details about the features, see sections of this manual). Pitch extraction robustly only works for single-line melodies. As the collection at this stage has not enough single-line melodies in all collections extracting and learning tonal systems, melodies, or melisma is left to the offline version. Below, on the very left top menu 'SOM Timbre Track u-Matrix' is displayed:

A trained SOM consists of neurons in a two-dimensional field. Each neuron is a vector of all features used for training, timbre, or rhythm features. Neigbhouring neurons might be more or less similar. This similarity is displayed in the u-matrix, where similar regions appear dark blue and blue, and more dissimilar regions appear green or yellow. Therefore we know that pieces in a dark blue region are very similar to one another, while pieces separated by a lighter ridge are more dissimilar.

The trained SOM is then used to place musical pieces on it, where each piece is placed at the neuron, which is most similar to this piece. Therefore, the trained SOM can be used to analyze pieces the map was not trained by, so e.g., new uploaded pieces from users.

The training set, here the ESRA pieces allow sorting pieces according to the musical content in this collection. Therefore the SOM is like a person knowing all these songs, but no others. For analysis of other musical styles or regions, the training of a new map might be considered. This can be performed in the offline version.

Therefore, the similarity or dissimilarity of a piece is a combination of two factors: a) the distance on the map between the two pieces and b) the coloring between the two pieces.

Exploring the map in terms of analyzing musical pieces in the online version is therefore done by exploring neighboring pieces. This can be done in several ways:

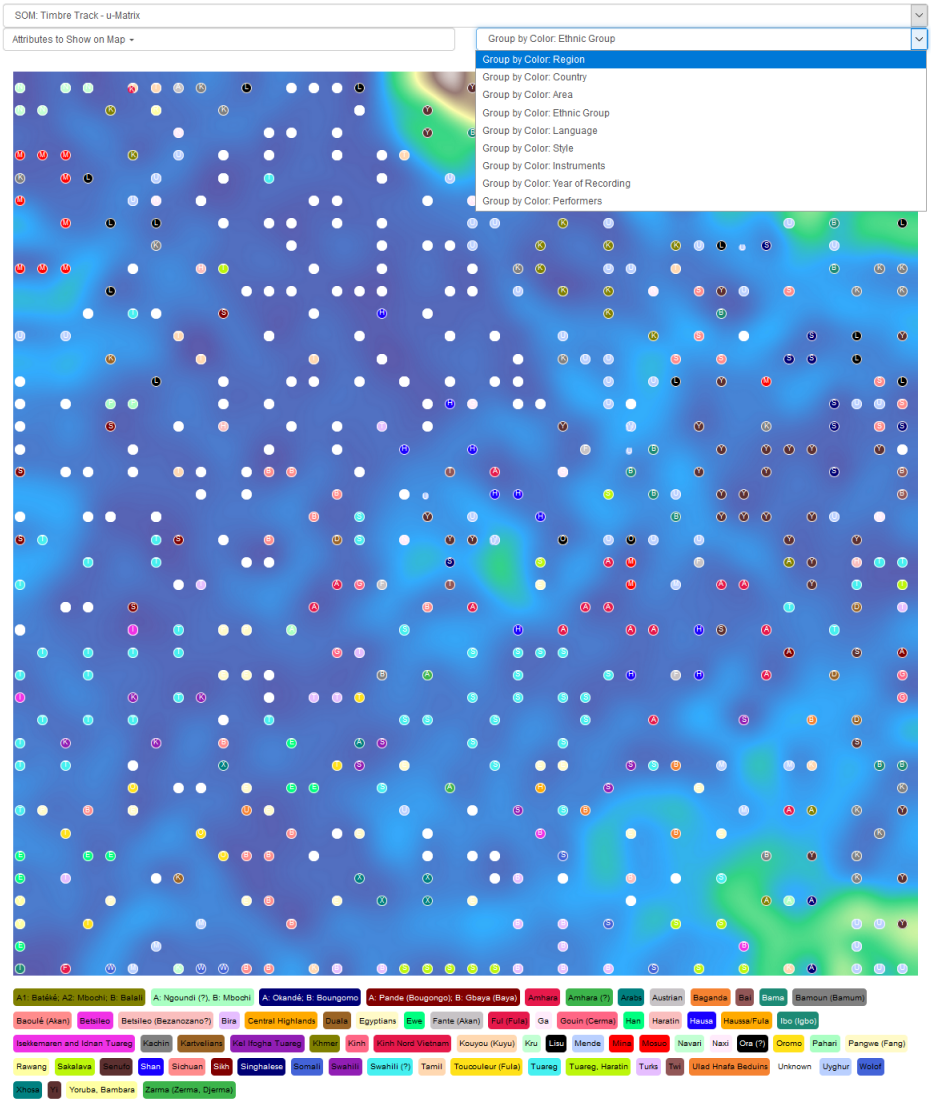
In the figure above, the pieces have colors according to ethnic groups, which is displayed at the menu entry on the top right. At the bottom of the plot, a legend shows the association between ethnic groups and color. The pop-up menu shows additional metadata. Choosing one of them changes the legend and the coloring of the pieces, respectively.

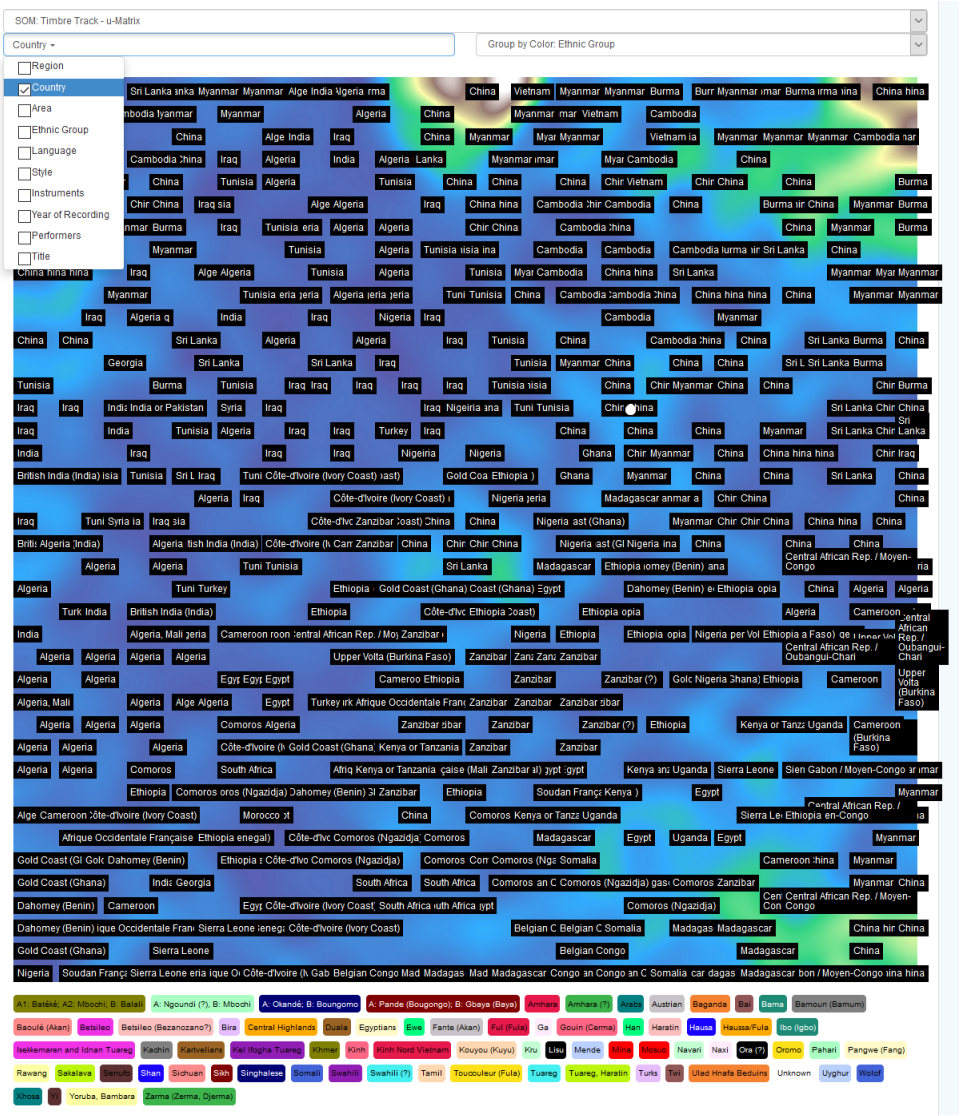
Another way of analyzing similarities is to display one or several metadata in the map. The top left menu allows several metadata to be used, also simultaneously. Be careful with this, the map might become too crowded:

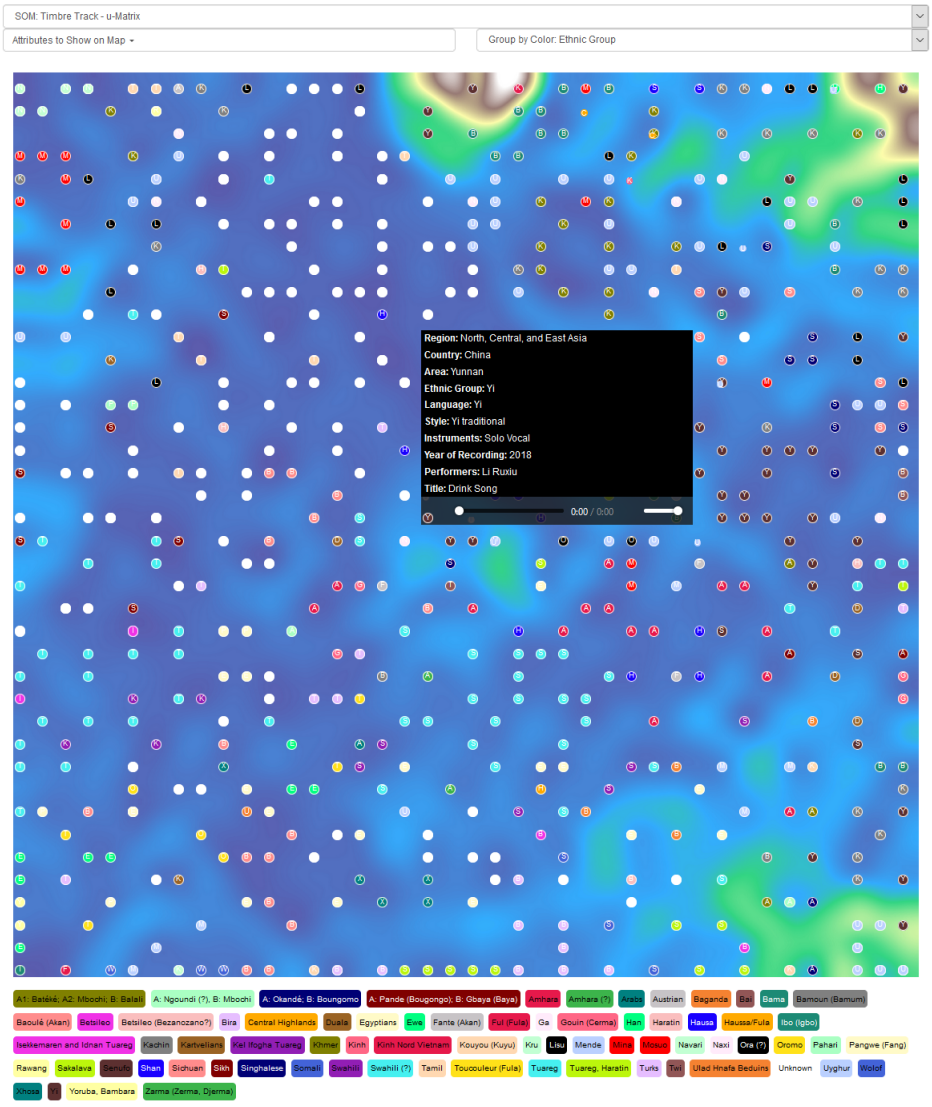
Yet another way is to look at the metadata of single pieces on the map. Moving the cursor over the map to a piece, the metadata of this piece pops up. Additionally, a player is shown, which allows instantaneous playback of this special song. It is very interesting to listen to neighboring pieces in this way, to hear if the analyzed feature of this piece fits aural perception.

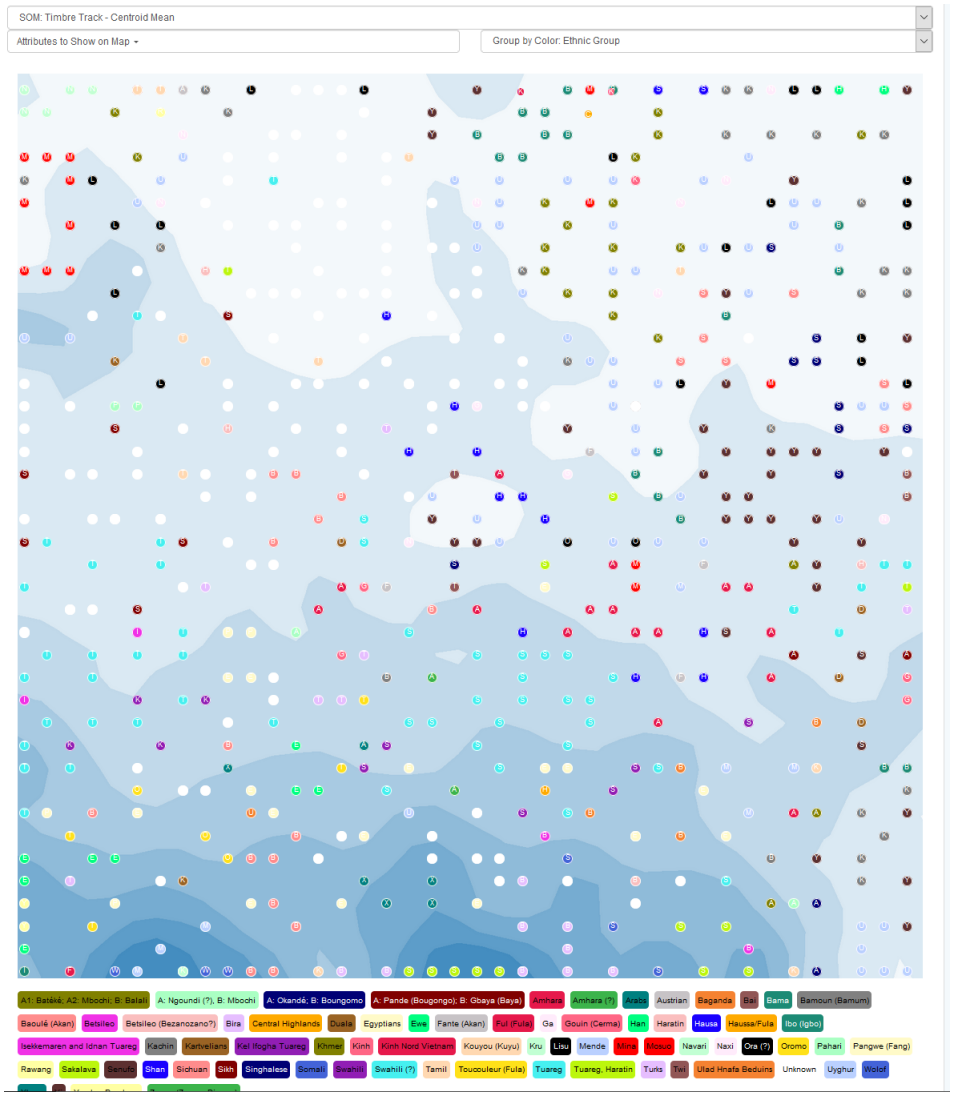
A fourth way of analyzing is to understand why the pieces are located in the map the way they are. Below examples of timbre are shown, rhythm is accordingly.

In the figure above, the top left menu displays 'SOM Timbre Track - Centroid Mean'. Then only the background image changes. In this case, it displays the strength of the spectral centroid, the perceived brightness of the songs. The


SOM: Timbre Track - u-Matrix
Attributes to Show on Map
Group by Color: Ethnic Group
Group by Color: Region
Group by Color: Country
Group by Color: Area
Group by Color: Ethnic Group
Group by Color: Language
Group by Color: Style
Group by Color: Instruments
Group by Color: Year of Recording
Group by Color: Performers
A1: Batéré; A2: Mbochi; B: Balali
A: Ngoundi (?), B: Mbochi
A: Okandé; B: Boungomo
A: Pande (Bougongo); B: Gbaya (Baya)
Amhara
Amhara (?)
Arabs
Austrian
Baganda
Bai
Baima
Bamoun (Bamum)
Baoulé (Akan)
Betsileo
Betsileo (Bezanozano?)
Bira
Central Highlands
Duala
Egyptians
Ewe
Fante (Akan)
Ful (Fula)
Ga
Gouin (Cerma)
Han
Haratin
Hausa
Hausse/Fula
Ibo (Igbo)
Isekkemaren and Idnan Tuareg
Kachin
Kartvelians
Kel Ifogha Tuareg
Khmer
Kinh
Kinh Nord Vietnam
Kouyou (Kuyu)
Kru
Liau
Mende
Mina
Mosuo
Navari
Naxi
Ora (?)
Oromo
Pahari
Pangwe (Fang)
Rawang
Sakalava
Senufo
Shan
Sichuan
Sikh
Singhalese
Somali
Swahili
Swahili (?)
Tamil
Toucouleur (Fula)
Tuareg
Tuareg, Haratin
Turks
Twi
Ulad Hnafa Beduins
Unknown
Uyghur
Wolof
Xhosa
Yoruba, Bambara
Zarma (Zerma, D)erma)

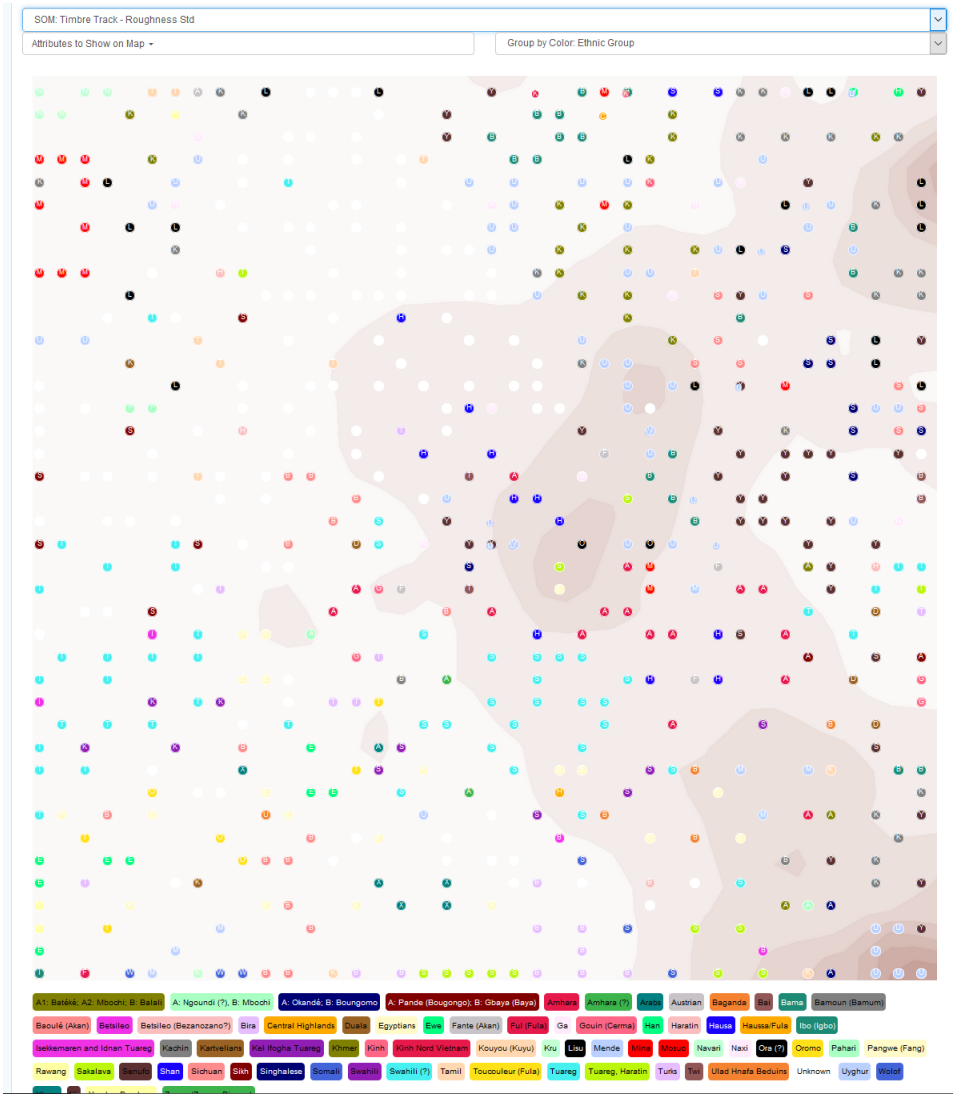

SOM: Timbre Track - u-Matrix
Country
Group by Color: Ethnic Group
Region
Country
Area
Ethnic Group
Language
Style
Instruments
Year of Recording
Performers
Title


SOM: Timbre Track - u-Matrix
Attributes to Show on Map
Group by Color: Ethnic Group
Region: North, Central, and East Asia
Country: China
Area: Yunnan
Ethnic Group: Yi
Language: Yi
Style: Yi traditional
Instruments: Solo Vocal
Year of Recording: 2018
Performers: Li Ruxiu
Title: Drink Song
0:00 / 0:00
A1: Bateké; A2: Mbochi; B: Balali
A: Ngoundi (?), B: Mbochi
A: Okandé; B: Bougomo
A: Pande (Bougongo); B: Gbaya (Baya)
Amhara
Amhara (?)
Arabs
Austrian
Baganda
Bai
Bama
Bamoun (Bamum)
Baoulé (Akan)
Betsileo
Betsileo (Bezanozano?)
Bira
Central Highlands
Duala
Egyptians
Ewe
Fante (Akan)
Fiji (Fula)
Ga
Gouin (Cerma)
Han
Haratin
Hausa
Hausa/Fula
Ibo (Igbo)
Iseikemaren and Idnan Tuareg
Kachin
Kartvelians
Kel Ifogha Tuareg
Khmer
Kinh
Kinh Nord Vietnam
Kouyou (Kuyu)
Kru
Lisu
Mende
Mina
Mosuo
Navari
Naxi
Ora (?)
Oromo
Pahari
Pangwe (Fang)
Rawang
Sakalava
Sanjhi
Shan
Sichuan
Sikh
Singhalese
Somali
Swahili
Swahili (?)
Tamil
Toucouleur (Fula)
Tuareg
Tuareg, Haratin
Turks
Twi
Ulad Hnafa Beduins
Unknown
Uyghur
Wolof
Xhosa
Yoruba, Bambara
Zerma (Zerma, Djerma)


SOM: Timbre Track - Centroid Mean
Attributes to Show on Map
Group by Color: Ethnic Group
A1: Batéké, A2: Mbochi, B: Bateli
A: Ngoundi (?), B: Mbochi
A: Okandé, B: Bougomo
A: Pande (Bougongo), B: Gbaya (Baya)
Amhara
Amhara (?)
Arabic
Austrian
Baganda
Bai
Bama
Bamoun (Bamum)
Baoulé (Akan)
Betsileo
Betsileo (Bezanozano?)
Bira
Central Highlands
Duala
Egyptians
Ewe
Fante (Akan)
Ful (Fula)
Ga
Gouin (Cerma)
Han
Haratin
Hausa
Hausa/Fula
Ibo (Igbo)
Isekkemaren and Idnan Tuareg
Kachin
Kantvelians
Kel Ifogha Tuareg
Khmer
Kinh
Kinh Nord Vietnam
Kouyou (Kuyu)
Kru
Lisu
Mende
Minia
Mosuo
Navari
Naxi
Ora (?)
Oromo
Pahari
Pangwe (Fang)
Rawang
Sakalava
Samutu
Shan
Sichuan
Sikh
Singhalese
Somali
Swahili
Swahili (?)
Tamil
Toucouleur (Fula)
Tuareg
Tuareg, Haratin
Turks
Twi
Ulad Hnafa Beduins
Unknown
Uyghur
Wolof

lower end of the two-dimensional plot shows dark blue regions. This means the centroid is much higher than in the other regions. Therefore, pieces placed here are much brighter, always compared to all other pieces on the map.

Another example below shows the standard deviation of roughness. This is achieved by changing the menu entry on the very left top. The pieces located on the right side have a much higher roughness standard deviation compared to the other pieces. This does not mean that they are rougher, it does mean that roughness over the course of the piece does change. The piece might be very rough at some point and very soft at another.



Combining all analysis tools allows for understanding similarities, dissimilarities, clustering, etc., of musical pieces in the collections and in songs uploaded by users. Basically, either one is looking for clustering in existing data, or one allows for exploring the pieces in terms of the analysis tools to start hearing similarities one might not be aware of before.

CHAPTER

SIX

ACCESSING YOUR OWN ARCHIVE WITH ESRA

6.1 Upload to ESRA

The easiest way to have your data analysed by ESRA is to upload it to the online archive. To do so, head to https://esra.fbkultur.uni-hamburg.de/ and log in. If you do not have an account yet, click on *Register* in the main navigation on the top of the page to creat a new account.

6.2 Create offline SOM

You can also create SOMs of your music collection on your own computer. ESRA is built upon apollon, an open source framework for audio feature extraction and music similarity estimation. A user friedly and easy to use abstraction of the apollon's functionallity is implemented in the comsar project. Please visit the comsar documentation for more information on installation and usage.

CHAPTER

SEVEN

REFERENCES

CHAPTER

EIGHT

CONTACT

8.1 Get in touch

We appreciate any kind of constructive feedback. If you wish to get in touch with us, please send an email to comsar[dot]ifsm[at]uni-hamburg[dot]de.

8.2 Feedback

No system is perfect. Did you find a bug in ESRA? Or do you have a feature in mind that would enhance ESRA? Please consider opening a discussion in our issue tracker.

If there is something wrong with this documentation that you like to see fixed, please report the problem in the documentation issue tracker.

- search

BIBLIOGRAPHY

[Bad13] Rolf Bader, editor. *Nonlinearities and Synchronization in Musical Acoustics and Music Psychology*. Volume 2 of Current Research in Systematic Musicology. Springer, 2013.

[Bad19] Rolf Bader, editor. *Computational Phonogram Archiving*. Volume 5 of Current Research in Systematic Musicology. Springer, 2019.

[Blass13] Michael Blaß. Timbre-based drum pattern classification using hidden markov models. In *Proceedings of the 6th International Workshop on Machine Learning and Music, ECML/PKDD*. 2013. URL: http://www.ecmlpkdd2013.org/wp-content/uploads/2013/09/MLMU_Blass.pdf.

[BlassFP20] Michael Blaß, Jost Leonhardt Fischer, and Nico Plath. Computational phonogram archiving. *Physics Today*, 73(12):50–55, 2020. URL: https://physicstoday.scitation.org/doi/10.1063/PT.3.4636.

[FZ07] Hugo Fastl and Eberhard Zwicker. *Psychoacoustics. Facts and Models*. Springer, 2007.

[SvRB09] Albrecht Schneider, Arne von Ruschkowski, and Rolf Bader. Klangliche rauhigkeit, ihre wahrnehmung und messung. In Rolf Bader, editor, *Musikalische Akustik, Neurokognition und Musikpsychologie*, volume 25 of Hamburger Jahrbuch fuer Musikwissenschaft, pages 101–144. 2009.

[Set93] William A. Sethares. Local consonance and the relationship between timbre and scale. *Journal of the Acoustical Society of America*, 94(3):1218–1228, 1993. URL: https://doi.org/10.1121/1.408175.

[Sot94] R. Sottek. Gehörgerechte rauhigkeitsberechnung. In *Fortschritte der Akustik: Plenar und Fachbeiträge der 20. Deutschen Jahrestagung für Akustik*, 1197–1200. DPG, 1994.

[vH63] Hermann von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg, Braunschweig, 1863.

[vR13] Arne von Ruschkowski. *Lautheit von Musik: eine empirische Untersuchung zum Einfluss von Organismusvariablen auf die Lautstärkewahrnehmung von Musik*. PhD thesis, Universität Hamburg, 2013. URL: https://katalogplus.sub.uni-hamburg.de/vufind/Record/78110422X?rank=1.