IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Music Content Selection Automation

Tetiana Gladkykh
Data Science Group, SoftServe
Lviv, Ukraine
thlad@softserveinc.com

Taras Hnot
Data Science Group, SoftServe
Lviv, Ukraine
thlad@softserveinc.com

Roman Grubnyk
Data Science Group, SoftServe
Lviv, Ukraine
thlad@softserveinc.com

***Abstract*— We are proposing the solution for musical content recommendation, which is based on assessment of tracks similarity with taking into account tree factors - genre description, sound and rhythm patterns and user preferences. We have introduced the music compositions distance measure based on their representation as mel-spectrograms, and deep-learning approach to high-level (tags) music description, based on the extracted acoustic and rhythmic patterns from their spectra.**

Keywords—music recommender; tracks similarity; mel-spectrogram; deep-learning; tags recognition

I. Introduction

If you look at a relatively close past, it becomes clear how much easier music lover life has become. Search of the right artist, composition or concert has become several orders of magnitude easier. Dozens of different services at our disposal that can satisfy the needs of even the most sophisticated music lover. Nevertheless, even with such music content variety, the problem of repertoire updating does not cease to be relevant - new performers working in the genre that is interesting right now, perhaps - in the mood; compositions that sound similar, have similar drums to something heard or liked. Existing services make the audio composition search directional and, certainly, ease the task of music content selection. They make it easier, but do not meet the challenge completely, because they work with a number of limitations. For example, the most popular service for music content search, Shazam [1, 2], uses prints of the spectra comparison as the core of the "search engine", so it turns out to be useless in case of, for example, a cover version of a famous work or new music content. Last.FM recommendation service [3, 4] allows each registered user profiling in order to perform his positioning regarding other registered participants and predict what else might be interesting for him based on the auditions general history. Among the weaknesses a fairly high secondary content percentage in the recommendations can be noted. The inability to recommend in the case of a new work and / or a new user. The Pandora service [4] based on comparison, which rests on the "contents" of a music piece evaluation that is expressed in the several hundred attributes set, provided by professional musicians. Yandex.Music [6], uses only user's listening history, it does not allow to segment songs by genre and directions, therefore, a fairly significant history of the user's activity is required to get adequate recommendations, and, as with two previous services, the problem of a "cold start" remains. If you take a whole galaxy of similar services, like TuneGlue, Music Roamer, Music-Map and others, they simply build a tree of compositions similarity, relying only on the musical works metadata, and can hardly be classified as a recommendation system.

We offer functionality that in many respects repeats existing services capabilities, in fact, combining them within a single product. But the distinctive feature of our solution is the search for works that sound alike, relying on the sound and rhythmic pattern, even if musical fragments don't match exactly.

II. Related Works and our contribution

In the context of music recommendation, we can mention three general approaches: 1) recommendation based on musical compositions metadata (like set of tags that describe musical genre, direction, artist, etc.), 2) recommendation based on context, like playlist, web-based co-occurrences, etc. 3) recommendation based on music feature extraction.

There are a lot of works are related to the first approach, it’s general advantage – relatively simple realization in the context of huge songs datasets and possibility to use users-based descriptions. But these methods require extremely detailed description in order to increase the results relevancy and can’t be applied in the case of new composition. The second approach allows to assess songs similarity based on the principle that two songs should be considered as similar if they are mentioned in the same context, so recommendations, in this case, may include so-called user-based similarity – similarity that is based on the user’s rates. The main advantage of this approach – the only information about song, that we should get, is the context, but this approach has significant restriction – we need some historical information about composition, so it also can’t be applied to some new song.

Recommendations that are based on music feature extraction also divided into two groups – high-level low-level features based recommendations. The low-level features [7] describe any audio signal in the form of well defined and determined acoustic features like: loudness, spectrum powers, brightness, bandwidth, pitch and cepstrum. The main disadvantage of these features – they can’t be easily used for understanding so-called structure of music to users without technical skills in this subject. This disadvantage is not specific for, so-called, high-level features – composite music characteristics like melody or harmony. This features describes the type of knowledge that a listener may extract, recognize and understand from one or other piece of music. There are works related to high-level feature extraction based on the chromagram analysis and estimation of the basic frequency corresponding to the pitch of the predominant melody with different modifications [8, 9]. All of these algorithms allow to extract complex music characteristics, based on the generalizations of the low-level music features processing, and representation them in more understandable form. So, high-level features may be considered as high-level interpretation of the set of low-level music features and, according to the recent works, low-level features are indispensable in the context of machine-learning approaches to music processing, understanding and tracks similarity assessment. For example, in work [10, 11] the similarity

978-1-5386-2874-4/18/$31.00 ©2018 IEEE

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

between spectrum are used for assessment similarity between corresponded audio-contents, in [12,13] the same representation was used for genres and artists' recognition.

In our approach we proposed algorithm of music compositions similarity assessment based on acoustic and rhythmic patterns that can be extracted from musical tracks' spectral representation. Moreover, we proposed deep-learning approach to high-level music description based on the the same initial representation. In addition to the above-mentioned, in our solution we combine several approaches to music similarity assessment – based on the songs' metadata and on the registered users' preferences analysis, so, finally we provide an opportunity to give recommendations by the extraction of the acoustic perception of the composition user like, which is supported by automatic identification of the genre, style and direction results. This, on the one hand, allows giving accurate estimates of whether a certain composition of previously unknown artist, will be liked by some registered user, taking into account their personal preferences. And, on the other hand, to select adequate content for new users by analyzing their audio library. Below we will consider the approaches underlying the musical works similarity evaluation based on the genre description, sound and rhythm pattern, registered users' preferences, automatic genre and stylistic affiliation of the music content determination, and the music content selection automation.

III. Similarity Assessment of Musical Compositions

To obtain a compositions list that can be recommended for listening to one or another user, a comprehensive method to evaluate similarity of musical works was developed. It includes a similarity assessment in three areas: 1) based on genre description - two compositions are considered close when they are described by a close set of genre tags; 2) based on sound and rhythm patterns - two compositions are considered close when they are characterized by close sound and rhythm patterns; 3) based on the registered users' preferences analysis - two songs are considered close when users with close preferences like them.

A. *Similarity evaluation based on genre description*

When we assess the compositions similarity, it is necessary to take into account many different factors, including the track description in the form of a tags set stamped by individual users. The more users provide this kind of characteristics, the more likely they match to track. Tags that are characterized by a large number of matches can be considered the most significant composition characteristics, since people with different, in most cases, preferences, were solidarity with their descriptions. The separate tag importance as a track's characteristic depends on the following factors: 1) ***popularity*** - depends on the number of users who in their track description indicated this characteristic; 2) ***uniqueness*** - a value indicating how this characteristic distinguishes a track against other tracks. So, the most popular tags are tags related to genres - general musical directions, such as rock, hip-hop, jazz, etc. Unique tags include, first of all, the performers' names, the album name, specific sub-genres, etc. For example, if we select compositions in which some genre direction is mentioned, $\text{track} = \{\langle \text{tag}_\text{i} \rangle\}$; $\langle \text{genre}_\text{j} \rangle \in \text{track}$, a composite tag can be considered as a sub-genre that contains the genre addition: $\langle \text{sub\_genre}_{\text{jk}} \rangle = \{\langle \text{genre}_\text{j} \rangle, \langle \text{description}_\text{k} \rangle\} \in \text{track}$.

Subgenre can be interpreted as a clarifying tag if it occurs much less frequently than the corresponding genre. So, if we consider 2 genre directions, we get the following frequency distribution for genres and sub-genres (Fig. 1).

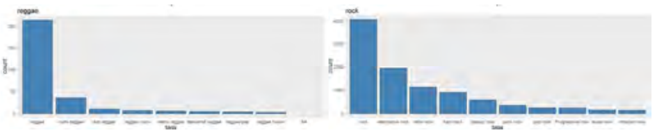


Fig. 1. Genres and Sub-genres distribution

You can see that subgenres may occur 4 times less frequently than the corresponding global directions, although such subgenres as "heavy metal", "alternative metal" and "alternative rock" can be considered as separate genres. Accordingly, when describing compositions, we can distinguish three tags' categories:

- General - genre affiliation. Tags are characterized by a high mention frequency in the context of a large compositions number
- Clarifying - subgenre affiliation. Tags are mentioned in a much smaller songs number than common tags, but they accompany the main genre tag
- Unique - artist or album characteristics. Tags are mentioned in the context of an extremely small tracks number, but each is one of the most popular particular work description

To rank the tags, we used a well-known statistical measure to evaluate the word importance in the context of a document that is part of the corpus - TF-IDF [14]. In our case, TF is the relative number of users who described a certain track with a certain tag, and IDF is a uniqueness measure, depending on the relative number of documents in which the tag was mentioned. Thus, each track can be described by the set VTrack of TF_IDF coefficients of tags mentioned at least once in the compositions' description. For this purpose, we used a bag-of-words model for documents presentation (in our case, tracks). Similarity between compositions S_z и S_y is estimated as the cosine distance between the corresponding vectors:

$$\text{Dist}_{\text{S}_\text{z}\text{S}_\text{y}} = 1 - \cos(\text{VTrack}_\text{z}, \text{VTrack}_\text{y}) \quad (1)$$

The proposed approach result is shown in the fig. 2. The figure shows a tag cloud describing the source composition – "India Arie – Beautiful" and one of the recommended songs – "Erykah Badu - Bad Lady". As you can see, the compositions perfectly correspond with each other in terms of genre and style descriptions.

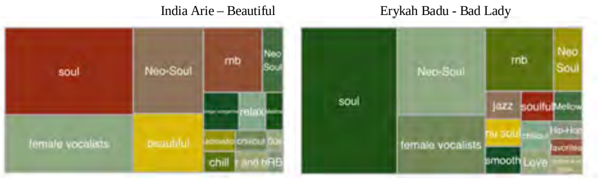


Fig. 2. Tags-based similarity

B. *Similarity evaluation by sound*

The second approach to similarity evaluation of compositions is based on the sound and rhythmic pattern formalization of individual tracks and their subsequent comparison.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Sound and rhythm patterns

At the core of the proposed solution is based on attempt to take into account the sound perception features by the organ of human hearing. This perception is estimated by a psychoacoustic value - the pitch, unit of which is “Mel”. By definition, "pitch" [15] is "the sound quality determined subjectively by a person using ear." Mel is an off-grid pitch unit, and to quantify system uses the results of data statistical processing on sound subjective perception [16-18]. The audio signal (sound) can be described by a set of so-called mel-cepstral coefficients - a representation of the spectrum power in the mel-frequency, obtained with separation of individual spectrum frequency range. To do this, windows that are evenly spaced on the mel-axis are used. To represent the music track, it is preliminary divided into fragments of short duration - about 23 ms, in order to describe the signal spectrum change character in time (the cases of signal non-stationary in the sections under consideration may be disregarded). The track fragment spectra combination makes it possible to describe the input signal in the form of a spectrogram, a two-dimensional function that displays the spectral power density of the signal dependence on time. In order to take into account the auditory perception peculiarities of sound to humans, we proceed to the mel-cepstral coefficients - the orthogonal logarithm mapping of the energy spectrum square at certain frequencies for a certain period of time:

$$S[n] = \sum_{m=0}^{M-1} P[m] \cos(\pi n\,(m + 0.5)/M)\,, 0 \le n < M$$

where,

$$P[m] = \ln\left(\sum_{k=0}^{N-1} |X_a[k]|^2\, H_m[k]\right), 0 \le m < M$$

$$X_a[k] = \sum_{k=0}^{N-1} x[n] e^{\frac{-2\pi i}{N} kn}\,, 0 \le k < N$$

$$H_m = \frac{(k - f[m-1])}{(f[m] - f[m-1])},\ \ f[m-1] \le k < f[m]$$

For each 23 ms fragment, we obtain a sequence of 40 cepstral coefficients, combination of which is a mel-spectrogram. The diagram (fig.3) shows visualizations of distinctive representatives of different musical genres.

Compositions similarity assessment

With a musical composition compact representation in which the person's auditory perception features are laid, we used it to assess the compositions similarity by sound. The core of the proposed method is the algorithm usage of the time scale dynamic transformation, which allows finding the optimal correspondence between time sequences.

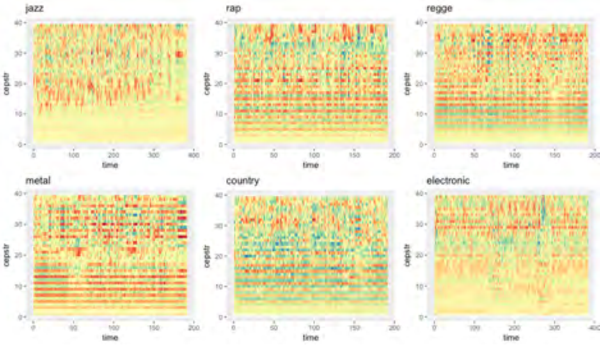


Fig. 3. Different genres spectrogramms

This algorithm is often used to compare time series and allows you to find a variant of their best alignment in order to level the error from an incorrect estimate of the distance associated with the possible rows displacement to each other. In its implementation, the algorithm is close to estimating the editorial distance when comparing two lines. The core of it is to build a distances matrix between all points pairs of the analyzed sequences A and B, after which the so-called transformation matrix is constructed:

$$D_{ij} = d_{ij} + \min\left(D_{i-1j}, D_{i-1j-1}, D_{ij-1}\right)$$

The distance between the sequences is the last element of the matrix. Number of points is m:

$$DTW(A, B) = D_{mm}$$

In our case ***time sequence*** - ordered in time set of mel-cepstral coefficients calculated for a fragment with a duration of 23 ms. Accordingly, each such set is interpreted as a ***sequence element***. The distance matrix is filled with cosine distances between the elements of 23 millisecond sequences of two compositions ***A*** and ***B***:

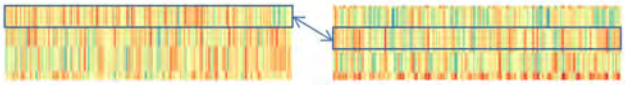


Fig. 4. Fragments comparison visuaization

$$d_{ij} = 1 + \cos\left(A_i, B_j\right)$$

Two compositions, represented by spectrograms *Sz* and *Sx*, are considered as close when they have quite a lot of matches in rhythmic pattern and sound. This means that we need to segment the tracks into sections that are long enough to catch this picture, but at the same time short enough to make the number of cases, when segment contains two or more sound patterns not too big. Each composition is characterized by a set of N “control” segments, therefore, their comparison consists in a complex estimation of the DTW distance between pairs of all control observations:

$$Dist_{S_z S_x}(N) = quantile_{25}\left(\left\{DTW_{r,k}\right\}\right)$$

where $DTW_{r,k} = DTW(S_{zr}, S_{xk})$, $r, k = \overline{1, N}$, N - number of control segments.

The optimal segment duration - 3 sec, was established empirically, by comparing clustering fragments results of 10 compositions with the corresponding classes (each composition is characterized by one class). It is expected that with properly estimated segment length, each of the resulting clusters will consist almost entirely of one composition fragments. The evaluation was carried out in accordance with the Rand index:

$$Rand = \frac{SS + DD}{DD + DD + SD + DD}$$

where SS – the number of elements pairs belonging to the same class and to one cluster, DD - the number of elements pairs belonging to different classes and different clusters, SD - the number of elements pairs belonging to the same class and to different clusters, DS - the number of elements pairs belonging to different classes and one cluster. Target optimization function in the context of segment length and criterion *Rand*: $\min_i \left(len_i, \frac{1}{Rand_i}\right)$.

The following figure shows an example of visualizing the t-SNE transformation of 10 different compositions. The compositions belong to different genres and are described by

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

a mel spectrograms combination for different window sizes (1 and 3 seconds, respectively).

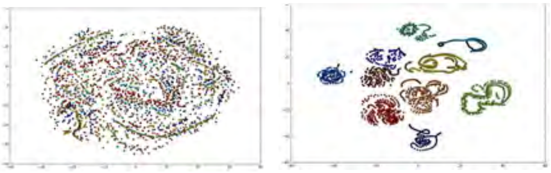


Fig. 5. T-SNE transformation of 10 different compositions

As can be seen, in the second case the compositions were divided into compact, well-divided groups, which confirms the result validity. The second parameter is the number of control fragments. The DTW distance between composition fragments estimating procedure is laborious enough - the computational complexity of O(N2), hence the computational complexity of estimating the distance between compositions is - O((M*N)2), by a factor of 1.5. The optimal number of segments was established on the basis of significance estimating of the discrepancy between the inter-composition distance for songs of one and different genres. The objective function should ensure the number N minimization of fragments with discrepancy maximization within and between genre distances:

$$\min_{\mathrm{N}} \left(\mathrm{N}, \mathrm{Dist}_{\mathrm{S_z S_x}}(\mathrm{N}) / \mathrm{Dist}_{\mathrm{S_z S_y}}(\mathrm{N})\right)$$

where $S_{z,}S_x$ - one genre, S_z, S_y - different genra

The following graphs show inside inner- and outer-genre distances, which depends on the number of tracks fragments:

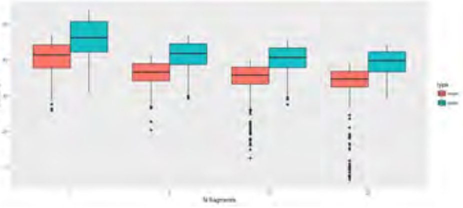
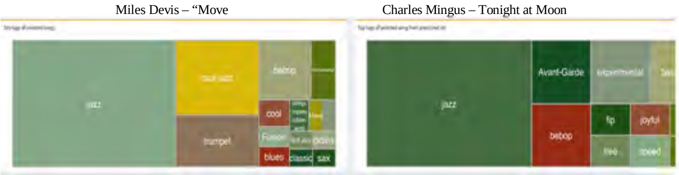


Fig. 6. Inner- and Outer- genre distances

As can be seen on the given diagram with the fragments number equal to 10, we get a satisfactory separation of the compositions inside and between the genre subgroups. With the number increase of picture fragments, the result improves slightly, accordingly number increase of segments is not justified. The similarity evaluation result of compositions by sound is shown in the Fig. 7. As in the previous example, the figure shows a tag cloud for the original song style and genre – Miles Devis – “Move” and one of the recommended songs – Charles Mingus – “Tonight at Moon”. As it is not difficult to see, the recommended composition belongs to the same direction as the original composition.



Miles Devis – “Move

Charles Mingus – Tonight at Moon

Fig. 7. Sound and rhythm based similarity

Similarity evaluation based on the preferences of registered users analysis

In addition to the objective characteristics of musical tracks, such as genre, rhythmic and sound patterns, which were discussed earlier, the closeness between individual compositions can be assessed on the basis of individual users’ subjective preferences. This technique is used in one of the methods, that is used in recommendation systems construction - item-to-item collaborative filtering, which is based on two objects similarity evaluation based on user estimates. Objects can be considered similar if they are liked by the same user group, or by users with similar preferences. In our case, we are dealing with a listening number matrix of users’ songs. The value of non-zero matrix elements can be interpreted as a kind of track rating according to the user - the higher the number of plays, the higher the confidence that the composition is included in the list of his preferences.

$$\mathbf{R} = (r_{ij}),\ \ i = \overline{1, \mathrm{N}_u},\ \ j = \overline{1, N_s},\ r_{ij} \geq 0$$

Because the plays number is not limited from above, unlike the objects rating, the matrix should be normalized with taking into account the user's activity. As the popularity rating of a song, we can correlate the plays number of a song with the upper bound of the corresponding $\alpha_q(\mathrm{R_i})$ distribution:

$$\mathrm{rate_{ij}} = \max\{\alpha_q(\mathrm{R_i}) | \mathrm{r_{ij}} \leq \alpha_q(\mathrm{R_i})\}, \alpha = 0.1n,\ n = \overline{0,1}$$

Since the tracks number is characterized by a histogram with right-hand asymmetry (*Sk*(listening)>1.5) (Fig.8, left), relation between rating and listening number can be represented by a logarithmic function (Fig.8, right). Parameters a and b depend on the user activity degree.

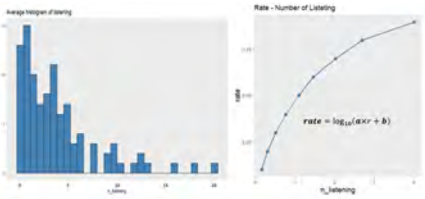


Fig. 8. Distribution of listening number and Rate function

Because the listening matrix is extremely sparse (about 1% of non-zero cells), the most optimal approach to collaborative filtering implementation is a model-based approach. As estimation model for users, a latent semantic model was chosen that allows to describe the system (user - composition) through a set of latent links. The solution is based on the listening matrix LU-factorization R′, which allows describing it as:

$$\underset{\mathrm{N_u \times N_s}}{\mathrm{R'}} = \underset{\mathrm{N_u \times d}}{\mathrm{L^T}} \times \underset{\mathrm{d \times N_s}}{\mathrm{U}}$$

where d – significant factors number, $\mathrm{U_j} \in \underset{\mathrm{d \times N_s}}{\mathrm{U}}$ vector of j composition in latent factors space, $\mathrm{L_i} \in \underset{\mathrm{N_u \times d}}{\mathrm{L}}$ vector of i user in latent factors space

Users and compositions representation in the latent factors space, their optimal number was evaluated based on the weighted estimates original matrix restoration error. The error weight is a non-decreasing function of the listening songs number and is given by the following expression:

$$\mathrm{w}(\mathrm{r_{ij}}) = \begin{cases} \alpha,\ \mathrm{r_{ij}} = 0 \\ \mathrm{r_{ij}^{\beta}}, \mathrm{r_{ij}} > 0 \end{cases}$$

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

where α – error weight with zero plays number, β- scaling factor

Optimization functions looks like this:

$$\sum_{ij}\left(w(r_{ij})(r'_{ij}-L_i^T\times U_j)^2\right)+\lambda\left(\sum_i\|L_i\|^2+\sum_j\|U_j\|^2\right)\rightarrow \min$$

With such a representation, the similarity estimation between two compositions S_z and S_y can be reduced to distance estimation between corresponding vectors in the latent factors space:

$$\mathrm{Dist}'''_{S_zS_y}=1-\cos(U_z,U_y)$$

The result of the proposed method is shown on Fig.9. As you can see, the recommended composition shows well the original work stylistics, and, consequently, it can be expected that the recommendation will be relevant to the user's preferences.

jazz
jazz piano
cool jazz
piano jazz
salsa
latin
Love
latin pop
male vocalists
italiana

Fig. 9. User preferences based similarity

IV. NEW COMPOSITION AUTOMATIC TAGGING

In advisory systems, the recommendation of new, previously unknown content, is particularly difficult, since there is often not enough information for its positioning among already existing objects. To solve this problem, we have developed a system of automatic musical work description (tagging), relying only on the analysis of its rhythmic and sound pattern. The solution is based on a model that allows the composition to be assigned to one of the 100 predefined stylistic classes based on the mel-spectrograms analysis of its fragments sets.

Stylistic classes are compositions collections that are close to each other in a genre descriptions combination expressed as a set of tags, and were obtained on the basis of more than 10.000 musical works in different genres and directions clustering. Clustering was performed using the method of agglomerate clustering with the optimal cluster number estimation using the Duda-Hart method. The closeness between the compositions is based on the genre description, and is given by (1). For example, the first class (cluster) consists of 288 works with the dominant genre of punk rock. The largest cluster contains 628 elements and is formed with compositions related to the set of R&B, POP and Hip-Hop. One of the small clusters (36 elements) refers to an alternative musical direction, uniting musicians from Iceland. The resulting clusters can be characterized by common (intersecting) tags, but their set is unique. Thus, each composition can be associated with one cluster (a unique set of describing tags) that reflects its genre and stylistic features and can be recognized on the basis of rhythmic and musical image analysis of this work.

The proposed solution is based on the recognition model of the genre and stylistic class by analyzing the mel-cepstral coefficients set, that describe each of the 3-second fragments of musical work. The general scheme of the automatic tagging system for the musical composition is shown in the Fig.10.

To model input $G_{genre}(x)$ of genre and stylistic recognition mel-spectrogram S_{Tx_i} is submitted. It describes 3 second fragment x_i of musical composition T. Model returns 100-component vector V_{Tx}, each element of which contains a confidence degree that the recognizable fragment belongs to one of 100 classes:

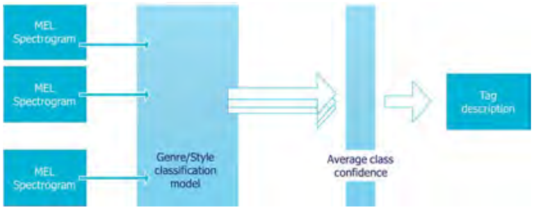


Fig. 10. General scheme of the automatic tagging system

$$V_{Tx}=G_{genre}(S_{Tx})$$

The composition genre is determined by averaging its fragments recognition results:

$$V_T=\frac{1}{n}\sum_{i=1}^{n}V_{Tx_i}$$

Final recognition of song's cluster is the set of top-n clusters, based of their confidence levels. The classifier model is based on an artificial neural network consisting of three convolutional and two fully connected layers. For model training we used 10`000 compositions of 30 sec lengths. Model accuracy was assessed like percent of correct clusters recognitions, when final recognition is the top-n clusters is equal to 65%, 72% and 86% for top-1, 2 and 3 clusters correspondingly. So, final set of tags is detected as set of n dominant tags from three clusters-winners:

$$\mathrm{Sel\_Tag}=\underset{n}{\mathrm{argmax}}\left(i|\left(\sum_{j=1}^{3}\sum_r\sum_i TF_{IDF_{irj}}\right)\right)$$

V. PLAYLIST COMPOSING

Approaches and functionality described above are the core of solution that allows to propose to user the list of compositions, which are relevant to their preferences, based on the analysis of several uploaded compositions. General flow as follows:

- Detection n songs, which are most close to the analyze composition / compositions based on rhythmic and sounds patterns (top-n (MFCC))
- For each song from top-n (MFCC): detection top-m songs, which are close by tags descriptions (top-m(Tags))
- For each song from top-n (MFCC): detection top-r songs, which are close by user preferences(top-r (Users))
- Play list consists of songs from all three lists – { top-n (MFCC)), top-m(Tags), top-r (Users)}

First stage is extremely time-consuming, because it requires assessment of MFCC distance between analyzed song and all songs from service database. In order to decrease the searching space, preliminary stage with detection the tags clusters is included to the flow. So, top-n MFCC distance based closest songs, are detected within songs from the clusters-winners (top 3 tags cluster). Final result is represented on the Fig. 11.

Analyzed song – "Love, love, love" by Monsters and Men was recognized like song that is characterized by following tags: Alternative Rock, Indie, Pop, Folk and Acoustic. Top 10 songs are represented on the figure (top-right part). This list –

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

is the list of the most relevant songs, songs that are close to the initial song by rhythmic and sound patterns. Below we can see additional songs, which were added based on tags and user based similarity. The most of songs belong to the artist from the first list, but there are some other compositions, which differ from initial song by sound, by very likely to be interesting to user.

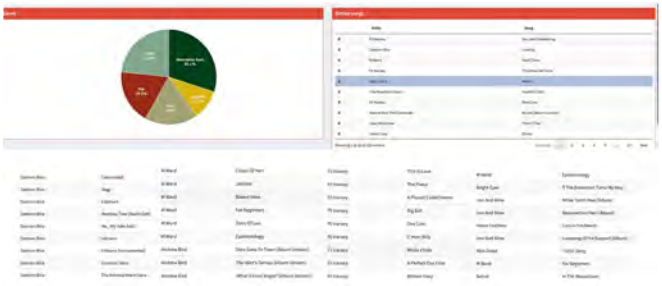


Fig. 11. Playlist composing

VI. CONCLUSIONS

In this work we have described the complex solution for musical content recommendation, which is based on tracks similarity assessment according to tree aspects: genre description, sound and rhythm patterns and the results of registered users' preferences analysis. We have proposed the distance measure between music compositions, which allows to assess the similarity between two and more tracks based on their representation as mel-spectrograms, and deep-learning approach to high-level (tags) music description, based on the extracted acoustic and rhythmic patterns from their spectra. Proposed solution allows to extract and describe user preferences based on perception of the compositions their like and provides the recommendations, which are supported by tags-based and user-based songs similarity. This, on the one hand, significantly improves recommendations that are based on users listening history and content-based tracks similarity only due to the possibility to put rhythmic and sound patterns of preferable compositions to the center of recommendations, but, on other hand, allows to solve so-called "cold start problem" – recommendation for new user, which doesn't have listening history or recommendations of new content (unknown genre and/or artist).

VII. REFERENCES

[1] "Shazam Launches Resonate TV Sales Platform," Billboard. 5 August 2014. Retrieved 15 June 2015.

[2] Bryan Jacobs, "How Shazam Works To Identify (Nearly) Every Song You Throw at It". Gizmodo. Retrieved 13 June 2017.

[3] Elia Alovisi, Last.fm: Was the Only Music Social Network That Made Sense, December, 2017, [https://noisey.vice.com/en_us/article/a37x9g/lastfm-was-the-only-music-social-network-that-made-sense]

[4] "Pandora and Last.fm: Nature vs. Nurture in Music Recommenders," Words & Numbers, A blog by Steve Krause, January, 2006 [http://blog.stevekrause.org/2006/01/pandora-and-lastfm-nature-vs-nurture-in.html]

[5] George Lawton, "How Pandora built a better recommendation engine," August 2017, [http://www.theserverside.com/feature/How-Pandora-built-a-better-recommendation-engine]

[6] Recommendation Technology 'Disco', [https://yandex.com/company/technologies/disco/]

[7] Florian Eyben, Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Springer, 2016.

[8] Justin Salamon, "Tonal Representations for Music Retrieval: From Version Identification to Query-by-Humming," International Journal of Multimedia Information Retrieval, vol. 2, iss. 1, pp 45–58, March 2013.

[9] J.,Salamon, and E. Gómez, "Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, iss. 6, pp 1759-1770, 2012.

[10] E. Allamanche and B. Froba, "Content-based identification of audio material using mpeg-7 low level description," In in Proc. of the Int. Symp. of Music Information Retrieval, pp 197–204, 2001.

[11] J. Wood and J. Dykes, "Spatially ordered treemaps," IEEE Transactions on Visualization and Computer Graphics, vol. 14(6), pp 1348–1355, 2008.

[12] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?,' In Proc. Int. Conf. Music Information Retrieval (ISMIR), Paris, pp. 157-163, 2002

[13] B. Logan and A. Salomon," A music similarity function based on signal analysis," In Multimedia and Expo,2001. ICME 2001. IEEE International Conference on, pp. 745–748, 2001.

[14] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press. 2008.

[15] Anssi Klapuri, "Introduction to Music Transcription," in Signal Processing Methods for Music Transcription, edited by Anssi Klapuri and Manuel Davy, New York: Springer, 2006, pp. 1–20. ISBN 978-0-387-30667-4.

[16] Stanley Smith Stevens, John Volkman and Edwin Newman, "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America, vol. 8 (3), pp 185–190,1937.

[17] Douglas O'Shaughnessy, Speech communication: human and machine. Addison-Wesley,1987. ISBN 978-0-201-16520-3.

[18] W. Dixon Ward, "Musical Perception," In Jerry V. Tobias. Foundations of Modern Auditory Theory. 1. Academic Press. 1970, pp. 405-447

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua