ARCHIVES OF ACOUSTICS
36, 2, 297–309 (2011)

DOI: 10.2478/v10168-011-0023-y

Automatic Music Summarization. A "Thumbnail" Approach

Jakub GŁACZYŃSKI, Ewa ŁUKASIK

Poznań University of Technology
Faculty of Computing Science
Institute of Computing Science
Piotrowo 3A, 60-965 Poznań, Poland
e-mail: Ewa.Lukasik@cs.put.poznan.pl

(*received November 27, 2009; accepted February 7, 2011*)

In the paper, various approaches to automatic music audio summarization are discussed. The project described in detail, is the realization of a method for extracting a music thumbnail – a fragment of continuous music of a given duration time that is most similar to the entire music piece. The results of subjective assessment of the thumbnail choice are presented, where four parameters have been taken into account: clarity (representation of the essence of the piece of music), conciseness (the motifs are not repeated in the summary), coherence of music structure, and overall quality of summary usefulness.

Keywords: automatic music summarization, music thumbnail, key phrases, automatic music transcription, Music Information Retrieval.

1. Introduction

Expansion of digital sound recording resulted in a sharp increase in the size of digital music collections. Easy access of music over the Internet, portable music players with a high storage capacity and the overall technological development mean that efficient management of extensive collections of musical tracks becomes a problem for large organizations and a growing number of individual music collectors around the world. Effective browsing of online music stores or music gathered on private computers as well as indexing, retrieval and management of tracks stored in large databases is easier if music summaries are applied.

As a summary of a piece of music we understand its fragment or a compilation of its fragments that is most representative, and that best reflects its essence, or includes its so-called Leitmotifs (key phrases). Such a definition of a summary may sound as a broad one and somewhat ambiguous, but it is largely dependent on a form of a given composition. Automatic music summarization is one of the problems of the domain of Music Information Retrieval (Kostek, Kania, 2009), which is in the scope of interest of the authors (e.g. Anioła, Łukasik, 2008; Dropik, Łukasik, 2010; Łukasik, 2005; 2010).

The research devoted to generate summaries of musical works have been carried out for more than a decade and related symbolic and audio domains. Most approaches are based on the discovery of repeated patterns in a musical work (Clifford *et al.*, 2006; Dannenberg, Hu, 2002; Xu *et al.*, 2009). This task may seem easier for symbolic representations, where semantic information is available (keys, pitch and length of notes, tempo etc.), then for audio domain, where all information has to be extracted from audio recordings. Meredith *et al.* (2002) proposed an algorithm for finding variations on a query pattern or for discovering patterns in polyphonic music, in which the voice of each note is unknown, using multidimensional point sets instead of strings.

In practice, the level of difficulty to automatically determine which parts of the piece of music best represent the whole work is different. It depends on the structure and characteristics of a piece of music. Often only selected patterns discovered by the algorithms are interesting for the listener and heuristics are proposed to identify patterns corresponding to themes, motives and other memorable musical patterns. Alternatively music patterns may repeat but each repetition may have varying structural characteristics, e.g. tempo, key, instrumentation, transformation etc. and sophisticated algorithms have to be used to discover them. The approaches to automatic music summarization will also differ for monophonic and polyphonic music. Moreover, music is often understood differently by different people, so the quality assessment of a summary is largely dependent on the subjective feelings of the evaluator. For professional musicians the exclusion of certain passages, even not repeated, may be unacceptable, due to e.g. a specific innovative harmony used by the composer. A non-professional listener will be more interested in fragments, which enable him or her to identify the piece of music in a short time, such as a chorus or a verse of a popular song.

In the case of popular music, there is a great deal of repetition of musical themes and the structure of musical piece is not complicated. Usually it is based on the schema ABACAB, where A is a verse, B – a chorus, and C – middle bars, called the bridge (Cooper, Foote, 2002). This makes the process of automatic music summarization easier.

The goal of the project described in the paper was to build a simple system for automatic summarization of pieces of popular music in audio domain. It is assumed, that musical pieces have repeated parts and a monophonic melodic line. The summary in this case is a continuous fragment of a given length that is most similar to the entire composition. No prior semantic information about the musical piece is needed (a pitch, a rythm, a key, etc.). Such a summary is often called "music thumbnail" (Bartsch, Wakefield, 2001; Kelly *et al.*, 2010). The algorithm consists of the extraction of parameters characterizing spectrum of a musical piece, choosing the similarity measure that will be used for calculating the similarity of the musical signal frames, constructing the self-similarity matrix and applying the aggregated similarity measure to find a starting point of a thumbnail. To validate the method, a thumbnails of a set of pieces of popular music were extracted and then assessed by a group of people during the individual listening tests. The system can be used in home collections to facilitate music retrieval.

The paper has the following structure: Sec. 2 discusses the methods of automatic music summarization, Sec. 3 introduces the method of a thumbnail extraction, Sec. 4 presents the experimental evaluation of the method, Sec. 5 discusses the outcome of the subjective assessment of the thumbnail choice and Sec. 6 concludes the paper.

2. Methods of automatic music audio summarization

A great deal of research has been devoted to automatic music summarization techniques in audio domain in the last decade. Some methods are based directly on the signal analysis, others first extract semantic information, e.g. pitch, chords, key, tempo, genre etc. (Chai, 2006). Two main categories of audio summaries defined by Peeters *et al.* (2002) are: "sequences" and "states" approaches, or a combination of those two. In the first approach the summary is a continuous excerpt that is most similar to the piece as a whole, often called a "thumbnail" (Bartsch, Wakefield, 2001). It is usually the most repeated component, thus the most representative part of a musical piece. The second approach first discovers the structure of a musical piece, identifies its characteristic parts (e.g. verses, choruses and the bridge) and then combines portions of each segment into one representative piece. Listeners' acceptance of such summaries is based not only on the selected fragments, but also on the smoothness of the transitions between them. The audio quality that is mostly used is timbre.

Foote (1999) introduced a similarity matrix to find frequent patterns in an audio representation of a music piece. Since then many methods have included feature based similarity matching for automatic music audio summarization. Examples of automatic music audio summarization methods are given e.g. by Chai (2006), Logan and Chou (2000), Peeters *et al.* (2002). Bartsch and Wakefield (2001) as well as Goto (2003) used pitch sensitive chroma-based features to detect repeated sections. Cooper and Foote (2002) defined a global similarity function based on extracted mel-frequency cepstral coefficients (MFCC) to find the most significant sections in the musical work. Logan and Chou (2000) used clustering and hidden Markov model (HMM) to detect the key phrases in the choruses. Xu *et al.* (2005) used an adaptive clustering method based on the linear prediction coefficients (LPC) and MFCCs together with SVM based adaptive clustering to automatic music audio summarization. Yang (2001) used dynamic time warping for the repeated parts of music with variations (e.g. tempo change). Xu *et al.* (2005) noticed, that human experts usually include in summaries transitions and small parts of verses appearing before and after a chorus, apart from the chorus itself. They proposed an automatic method, which mimics such an approach. The meaningful segments are extracted from verse, chorus and transitions between them.

Many authors perform experiments with users asking them to assess the quality of the automatically extracted summaries. User satisfaction is the most significant criterion of the quality of music summary.

3. Music audio summary as a fixed-length continuous segment

3.1. Introductory remarks

The approach described in this paper is "thumbnailing", i.e., according to Peeters *et al.* (2002), finding a fixed-length continuous music segment most similar to the entire song. Duration of the segment (the thumbnail), that will be denoted L, is arbitrary, in practice often 30 sec. The method is applicable to pieces of popular music with repeatable parts. Usually if we can identify the repeated sections in a piece of popular music, we are likely to have also identified a good thumbnail. The algorithm evaluated in this paper is based on the method presented by Cooper and Foote (2002) and is composed of the following steps:

- music signal parametrization,
- calculation of the distance metric value,
- construction of the similarity matrix,
- calculation of the aggregated measure of similarity,
- extraction of the "thumbnail".

The following subsections will present the details of these steps.

3.2. Signal parametrization

The music audio signal for algorithm development was uncompressed CD audio recording and decompressed mp3 file decoded to .wav format. Audio waveform was resampled to the sampling frequency 16 kHz, with right and left channels mixed to mono format. The frame length for the analysis was 100 ms and Hamming window without overlapping was used. The parameters calculated in the first step of the algorithm characterized spectrum envelope of the music audio signal. These were simple to extract, efficient in various application of audio analysis:

- 13 first mel-frequency cepstral coefficients (MFCC),
- the spectral centre of gravity that is attributed to the sound "brightness".

Feature vectors were standardized using Z-score:

$$z = \frac{x-\mu}{\sigma}, \tag{1}$$

where x is a random variable (feature value), μ – its mean, σ – standard deviation.

3.3. Distance measure

The structure of the musical piece may be found using a 2-D similarity or distance matrix of consecutive frames. Distance is bigger when two feature vectors are more different. Conversely the similarity gets the peak value if two feature vectors are identical. The distance d may be transformed into similarity s, using formulas dependent on the distance (similarity) measure, e.g.:

$$s = -d \tag{2}$$

assuming that $d \neq 0$, or

$$s = \exp(-d). \tag{3}$$

Choosing the appropriate measure of distance or similarity is one of the essential elements for discovering structure of a musical piece. Some of them have very similar behaviors in similarity queries, while others may behave quite differently (Quian *et al.*, 2004). The most popular is the *Euclidean distance*. It is calculated as the root of square differences between coordinates of a pair of objects:

$$d_e(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{\sum_{k=1}^{K} (\mathbf{v}_i[k] - \mathbf{v}_j[k])^2}, \qquad k = 1, \ldots, K, \tag{4}$$

where $\mathbf{v}_i$, $\mathbf{v}_j$ are vectors of parameters for frames i and j, K – dimension of parameters vector, $\mathbf{v}_i[k]$ is k-th coordinate of a vector $\mathbf{v}_i$.

The *cosine distance* gives the angular cosine distance between feature vectors, which is not sensitive to the energy of signals, as its values are naturally normalized. This measure is calculated as the scalar product of feature vectors, normalized by the product of their length. It represents the cosine of the angle between vectors of parameters in the K-dimensional feature space, where K is the number of features:

$$d_{\cos}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}, \tag{5}$$

where $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ is the scalar product of vectors in K-dimensional parameters space, $\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|$ is the product of norms of vectors.

Another similarity measure is derived from the *Tchebychev distance*, where the distance between two vectors is the greatest of their differences along any coordinate. Thus it is very sensitive to the differences between the values of coordinates of vectors:

$$d_{ch}\left(\mathbf{v}_i, \mathbf{v}_j\right) = \max_{l} \left|\mathbf{v}_i[k] - \mathbf{v}_j[k]\right|, \qquad k = 1, \ldots, K, \tag{6}$$

where $\mathbf{v}_i$, $\mathbf{v}_j$ are vectors of parameters, K – dimension of parameters vector, $\mathbf{v}_i[k]$ is k-th coordinate of a vector $\mathbf{v}_i$.

3.4. Construction of the similarity matrix and a thumbnail extraction

Once the similarity or distance between all possible frames (represented by their feature vectors) are calculated, a 2-D similarity (or distance) matrix is constructed with the elements $\mathbf{S}(i,j)$ equal to the similarity value of vectors $\mathbf{v}_i$ and $\mathbf{v}_j$:

$$\mathbf{S}\left(i,j\right) = s\left(\mathbf{v}_i, \mathbf{v}_j\right).$$

In this way the degree of the similarity of i-th frame to all other frames of the entire audio piece is denoted and the structure of the musical piece may be discovered.

The music summarization algorithm has to find the excerpt of a given duration, that is most similar to the entire piece of music. This is because the song usually has several repeated segments, e.g. chorus. The principle of the method is to find the aggregate similarity between the excerpt of the required duration L and the entire song by summing up columns (or rows) of similarity matrix for the appropriate number of frames, starting from different consecutive frames (initial points of an excerpt). The largest score indicates the segment that represents in the best way the entire piece of music. The score may be normalized by the length of the segment, so that the summaries of various lengths may be compared. This method was introduced by Cooper and Foote (2002).

To find the optimal summary of length L one has to find the excerpt of that length with the maximum summary score. Let us define that score as $Q_L(i)$:

$$Q_L(i) = \overline{\mathbf{S}}(i, i{+}L) = \frac{1}{NL}\sum_{m=i}^{i+L}\sum_{n=1}^{N}\mathbf{S}(m,n) \quad \text{for} \quad i = 1,\ldots,N{-}L, \tag{7}$$

where N is the number of frames in the entire song. The index of a summary starting frame that maximizes the summary score is:

$$q_L^* = \arg\max_{1\le i\le N-L} Q_L(i). \tag{8}$$

The maximum summary score q_L^* indicates the starting point of a thumbnail and $q_L^* + L$ is its ending point. If it happens, that two excerpts have the same value $Q_L(i)$, the earlier one is taken as a summary.

4. Experimental evaluation of the method

4.1. Thumbnail extraction

As it was stated in Subsec. 3.1 – the set of parameters and the distance measure affect the quality of the extracted thumbnail. Cepstral coefficients and brightness are very simple to extract and robust in various audio signal applications. The choice of the distance measure was preceded by tests in which Tchebyshev gave the most satisfactory results.

Let us analyze an example illustrated in Figs. 1–3. Figure 1a represents a similarity map for Queen's "Show must go on". A similarity map represents each similarity value by a pixel of a given value from the grayscale. This map allows us to identify the structure of the song. It was constructed using 13 MFCCs, cosine distance and the similarity of $L = 30$ s long segment with a starting point sliding frame by frame. As it was stated in Subsec. 3.1, the duration of the segment L, i.e. the thumbnail, is arbitrary, but most often 30 seconds.

From the analysis of Fig. 1b where the aggregated similarity of $L = 30$ s long fragment is presented as a function of a starting point, the optimal fragment to start would be around 150th second. Precisely, the index of a summary starting

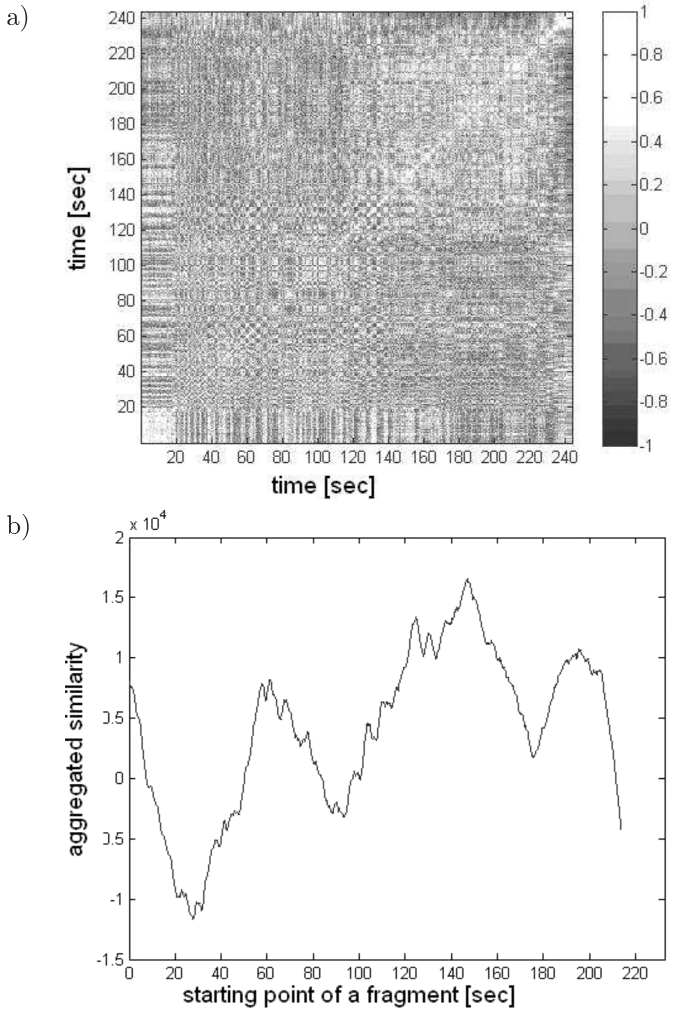

Fig. 1. Queen's – "Show must go on": a) similarity map using MFCC, center of gravity and cosine similarity measure, b) aggregated similarity of a 30 s long fragment as a function of a starting point of a fragment. Maximum is in 146th second.

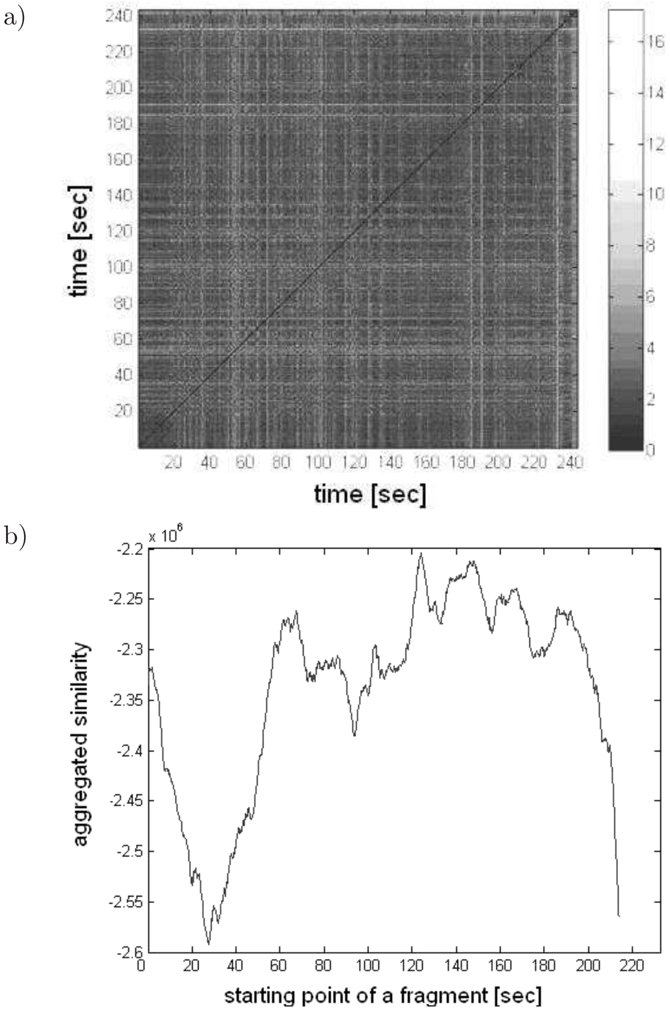

Fig. 2. Queen's – "Show must go on": a) similarity map using MFCC, spectrum centre of gravity and Tchebyshev distance measure, b) aggregated similarity calculated from Tchebyshev distance using formula (2) of a 30 s long fragment as a function of a starting point of a fragment. Maximum is in 123rd second.

frame maximizing the summary score q_L^*, calculated using formula (8), indicates 146th second. However the excerpt starting in this moment would include a guitar solo from 149th to 169th second – the bridge. The listener would not accept this fragment as the best candidate for the thumbnail of this song.

Let us now analyze Fig. 2a representing the aggregated similarity map obtained for a sliding window using Tchebychev similarity measure (negative of the Tchebyshev distance (6)) calculated for 13 MFCCs and a centre of gravity as features. As may be seen in Fig. 2b, the optimal summary starting frame q_L^*, is in 123th second. The excerpt of the length $L = 30$ seconds that starts at this point contains a full chorus of a song from 128th to 148th second. Therefore it would be a better recognizable fragment, then the previous one and could play the role of the thumbnail.

Figure 3 represents the time envelope of the analyzed song, where the starting points of the chorus repetitions and the solo guitar have been marked, as well as a 30 second long thumbnail (found using the Tchebyshev distance).

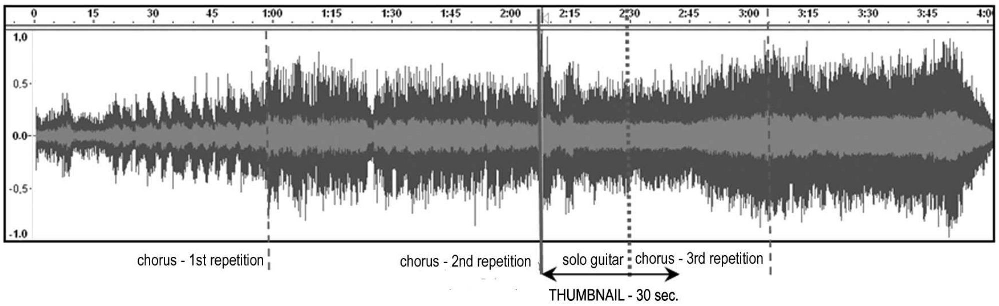

Fig. 3. Queen's – "Show must go on" waveform envelope. Starting points of the chorus repetitions, solo guitar and the thumbnail are indicated. The thumbnail starting point was calculated using Thebyshev distance.

Pieces of music, that constituted the subject of experimental evaluation are ten popular English and Polish songs and two popular classical music works of W.A. Mozart with a highly repetitive construction: Rondo alla Turca (3rd part of the Piano Sonata No. 11 in A major, K 331) and Dies Irae (introductory part of the Sequence in Requiem Mass in D minor, K 626). The time boundaries of 30 second long thumbnails are presented in Table 1. All of them have been

Table 1. The quality of summary assessment results for ten test songs.

Performer – Title	Summary time boundaries		Average assessment values			
	Start [sec]	End [sec]	K	T	L	O
50 Cent – In da club	82	112	4.9	4.2	4.3	4.6
Akcent – Tabu tibu	30	60	4.0	2.6	3.7	3.3
ATB – Let U Go	84	114	4.1	3.4	4.1	3.6
Human League – Together in Electric Dreams	190	220	4.3	3.7	4.3	4.2
Queen – Show must go on	123	153	4.8	4.6	4.0	4.7
Rihanna feat. Jay-Z – Umbrella	204	234	4.7	4.0	4.2	4.3
SDM – Jest już za późno, nie jest za późno	99	129	4.9	4.5	4.7	4.8
Sława Przybylska – Już nigdy	165	195	4.8	4.8	4.2	4.7
W.A. Mozart – Rondo alla Turca	161	191	3.9	4.0	4.0	3.9
W.A. Mozart – Requiem (Dies irae)	2	32	4.9	4.9	4.7	4.9
Total			4.53	4.07	4.22	4.30

calculated from the same parameters and Tchebyshev distance measure. For the compactness of presentation and the easiness of the analysis, they are collected together with the subjective assessments, commented in the following section.

5. Subjective assessment of the thumbnail quality

5.1. Assessment criteria and setting of the experiment

To subjectively assess the suitability of music thumbnails extracted in the way described in Subsec. 4.1. a set of criteria was needed. Some authors compare a thumbnail extracted automatically with the one extracted manually, others compare it with the excerpt chosen randomly. We have adopted assessment criteria used by Xu *et al.* (2005). These are: clarity, conciseness, coherence and overall quality of chosen excerpt. They answer the following questions:

- Clarity (K) – does the thumbnail represent the essence of a song?
- Conciseness (T) – does the thumbnail minimize repetitions of the same motifs?
- Coherence (L) – is the thumbnail reasonably chosen?
- Overall quality (O) – is the thumbnail representation satisfactory for the listener?

The ranking scale spans from 1 to 5 that corresponds to the worst and the best marks respectively.

5.2. Listening experiment results

The method of a thumbnail generation presented in this paper is related to popular music with a clear and simple structure. Listening tests were conducted with evaluators, who liked listening to the popular music and were not professional musicians. Neither their ability to understand musical coherence in relation to music theory, nor other musical skillfulness (e.g. short term memory or long term memory operating with various forms of pitch as discussed by Rakowski, (2009)) was tested – the evaluators represented average users of music digital repositories.

Ten volunteers took part in the experiment. They were not familiar with the idea of automatic music audio thumbnail generation. They could listen to the entire piece and to the thumbnail as many times as they wished before the task completion. They could also attach descriptive comments to each questionnaire, which would be useful for further refinement of the method. The evaluators worked individually and assessed ten pieces of music, listed in Table 1. In the same table the averaged scores for individual criteria for each piece are presented. The graph in Fig. 4 presents the results for individual criteria averaged for all songs.

The general result of the experiment was good – the average score for all criteria was 4.28. The highest score was given to the clarity of choice (K) – 4.53

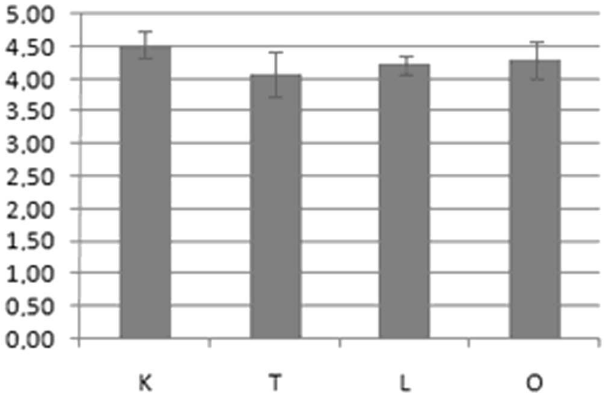

Fig. 4. Mean and standard deviation of summary quality assessment in various categories (K – clarity, T – conciseness, L – coherence, O – overall quality).

and for over-all quality – 4.30. Most listeners agreed in scoring the coherence of the summary. The conciseness was scored minimally for the song number 2 – “Tibu tabu”, where the repetition of a musical phrase occurred in the thumbnail. The best overall quality was assigned to Dies Irae of W.A. Mozart – 4.9. None of other pieces and criteria got such a high score. It is probably because of the clarity of the construction of this classical musical piece. It outperformed “Rondo alla Turca”, which has also a transparent, repeatable construction, but the fragment presented as the thumbnail came from the final part of the music piece whereas listeners generally preferred earlier occurrence of a characteristic motif.

A synchronization of the summary starting point with the beginning of a musical phrase appeared very important in accepting the summary. Even a very slight delay or precedence of the starting point in relation to the phrases were disapproved by listeners.

6. Conclusions

In the paper an algorithm for automatic music audio summarization has been presented. The music summary is understood as an excerpt of fixed duration, that is most similar to the entire music piece and constitutes its most representative part – “a thumbnail”. The similarity concerned features representing only spectral envelopes of audio signal frames – cepstral coefficients in mel scale and spectral center of gravity (brightness). Its calculation was based on Tchebyshev distance measure. A self-similarity matrix was constructed and the aggregated summary measure, introduced by Cooper and Foote (2002) was applied to find a starting point of a thumbnail.

The subjective assessment of a thumbnail choice has been carried out using four criteria: clarity of a summary, its conciseness, coherence of structure, and overall quality. The tests have shown users satisfaction with the extracted thumbnails (average score 4.28 and maximum score 4.9 in 1–5 scale). Descriptive comments from the evaluation confirmed the fact that different people expect different representation of the music. They also showed, that in certain cases additional excerpts from the composition would be required to fully characterize a musical piece.

References

1. Anioła P., Łukasik E. (2007), *JAVA library for automatic musical instruments recognition*, AES Convention Paper 7157.
2. Bartsch M.A., Wakefield G.H. (2001), *To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing*, Proc. WASPA.
3. Chai W. (2006), *Semantic segmentation and summarization of music: methods based on tonality and recurrent structure*, Signal Processing Magazine, IEEE, **23**, 2, 124–132.
4. Clifford R., Christodoulakis M., Crawford T., Meredith D., Wiggins G. (2006), *A fast, randomised, maximum subset matching algorithm for document-level music retrieval*, Proc. ISMIR.
5. Cooper M., Foote J. (2002), *Automatic Music Summarization via Similarity Analysis*, Proc. ISMIR.
6. Dannenberg R.B., Hu N. (2002), *Pattern discovery techniques for music audio*, Proc. ISMIR.
7. Dropik Ł., Łukasik E. (2010), *Two-Level Hierarchical Classification of Music Genre for Music Social Networks*, Foundations of Computer and Decision Sciences, **35**, 4.
8. Foote J. (1999), *Visualizing Music and Audio using Self-Similarity*, Proc. ACM Multimedia 99, pp. 77–80.
9. Goto M.A. (2003), *Chorus-Section Detecting Method for Musical Audio Signals*, Proc. IEEE ICASSP.
10. Kelly C., Gainza M., Dorran D., Coyle E. (2010), *Audio Thumbnail Generation of Irish Traditional Music*, Irish Systems and Signals Conference, Cork.
11. Kostek B., Kania Ł. (2008), *Music information analysis and retrieval techniques*, Archives of Acoustics, **33**, 4, 483–496.
12. Logan B., Chou S. (2000), *Music Summarization Using Key Phrases*, Proc. IEEE ICASSP.
13. Łukasik E. (2005), Wavelet *Packets Features Extraction and Selection for Discriminating Plucked Sounds of Violins*, Lecture Notes "Advances in Soft Computing", Springer-Verlag, pp. 867–875.
14. Łukasik E. (2010), *Long Term Cepstral Coefficients for violin identification*, 128 AES Convention Paper 8132, London.
15. Meredith D., Lernstrom K., Wiggins G.A. (2002), *Algorithms for discovering repeated patterns inmultidimensional representations of polyphonic music*, Journal of New Music Research, **31**, 4, 321–345.

16. Peeters G., Burthe A., Rodet X. (2002), *Toward automatic music audio summary generation from signal analysis*, Proc. ISMIR.

17. Rakowski A. (2009), *The domain of pitch in music*, Archives of Acoustics, **34**, 4, 429–443.

18. Xu Ch., Maddage N.C., Shao X. (2005), *Automatic Music Classification and Summarization*, IEEE Transactions on Speech and Audio Processing, **13**, 3, 441–450.

19. Xu J.P., Zhao Y., Chen Z. (2009), *Music snippet extraction via melody-based repeated pattern discovery*, Sci China Ser F-Inf Sci, **52**, 5, 804–812.

20. Yang C. (2001), *MACS: Music Audio Characteristic Sequence Indexing for Similarity Retrieval*, Proc. of WASPAA.