Protecting Copyright Ownership via Identification of Remastered Music in Radio Broadcasts

Pasindu Marasinghe, K.L. Jayaratne and M.I.E. Wickramasinghe
University of Colombo School of Computing, Sri Lanka

***Abstract*—Music identification in radio broadcasts have different use cases like playlist generation and protecting copyright ownership of artistes. In real world environments, music is often remastered by radio channels to fit in to limited air times. Mostly those remastering includes time stretching, pitch shifting and etc. Therefore robustness of the music identification method plays a crucial role. In this paper, we propose to use Scale Invarient Feature Transform (SIFT) on Short Time Fourier Transformation (STFT) spectrogram to extract audio descriptors. Experiments show that SIFT descriptors exhibit robustness against audio distortions such as time stretching and pitch shifting. Finally a ratio based threshold is used to differentiate identified and non-identified states.**

***Index Terms*—Audio Fingerprinting, Music Identification, Radio Broadcast Monitoring**

I. Introduction

According to the intellectual property act of Sri Lanka[1], royalties must be paid to the original artistes when a song is broadcast on a radio channel. Each radio channel is maintaining a playlist to keep track of the songs that were broadcast throughout the day. That playlist can later be used to pay royalties to the respective artistes. However, in order to streamline and regulate the royalty payment process, it is vital to have a method to monitor the radio broadcasts. Manual radio broadcast monitoring is infeasible and expensive due to increasing number of both radio channels and songs. In manual monitoring a person should be assigned to each channel who needs to keep record of each song in the radio broadcast of that assigned channel. Due to the increasing number of songs and the fallible nature of humans such a monitoring task is prone to errors and inaccuracies. Hence an automated radio broadcast monitoring approach must be considered as a viable alternative in the modern day radio broadcast monitoring.

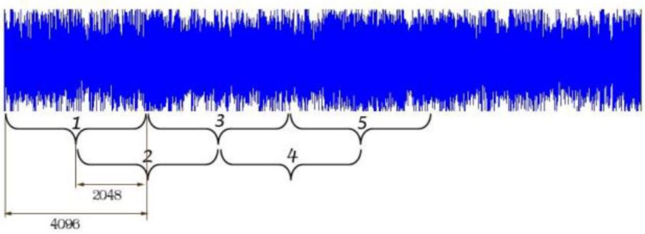

Fig. 1. Key controlling parameters of STFT[2]

In the research "Radio Broadcast Monitoring to Ensure Copyright Ownership"[2], researchers have implemented an automated radio broadcast monitoring system (refer the Figure 2 for the architecture) which has achieved 97.14% overall accuracy in identifying original songs in radio broadcasts. The researchers introduced an audio fingerprint to register and identify songs. The fingerprint was introduced as a series of hash values extracted from frequency domain audio signal. Time domain signals were converted to frequency domain by using STFT, which used 4096 bits long window and 2048 bits long overlapping area as shown in Figure 1. Then five peak values were extracted for each window by dividing mid frequency level into five bins and taking the peak value from each bin. Extracted five peak values were used to create a hash value as depicted in Figure 3.

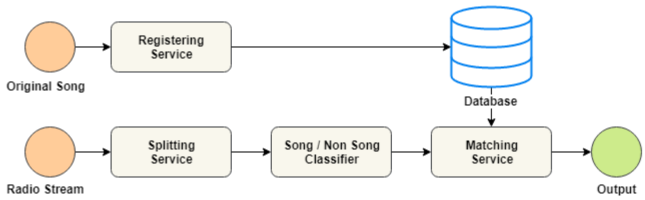

Fig. 2. Architecture of the existing system

In contemporary radio broadcasts, channels tends to alter songs by including commercials and dialogues and by remastering the original song. Remastering can be done by adding or subtracting elements, or by changing pitch, equalization, dynamics or tempo[3]. Even though the above mentioned radio broadcast monitoring system's accuracy is not significantly affected by commercials and dialogues included in songs, the system is unable to identify a song when that song is remastered by the radio channel as changing pitch, equalization, dynamics or tempo directly affects both time domain and frequency domain audio data.

Timbre, tempo, timing, structure, key, harmonization and lyrics are the basic musical facets that can be identified[3]. Timbre, also known as tone colour is the music facet which makes a difference of different sound productions even when they have the same pitch and loudness. Simply it is what makes a difference between a piano and a violin playing the same note at the same volume. Timbre can be changed due to the

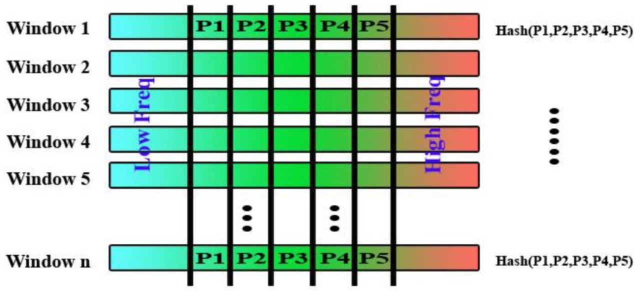

Fig. 3. Extracting peaks and generating a hash value[2]

use of different sound enhancing and processing techniques or to the use of different instruments and configurations. Tempo is the speed or pace of the music which can be easily changed by playing the music at different speeds. The music facet of timing is a rhythmic structure of the music which can be altered by the changes to the drum section. Structure is the arrangement of music sections, and music structure alterations can be made while remastering. Key, harmonization and lyrics are tonality, chords and words of the music which can be altered while remastering.

In order to identify remastered music in radio broadcasts, existing literature on cover song identification and music similarity measures can be used as foundation study to this research. Directly implementing a cover song identification method or a music similarity measure to identify remastered music in radio broadcasts is not possible as there is limited time to do the identification, because it is not just comparing two music clips to find similarity, but comparing a radio broadcast with a more than twenty thousand song database.

II. Related Work

Remastered song identification falls under the domain of cover song identification, which is a very active area of study in the Music Information Retrival (MIR) community [3]. Literature on cover song identification contains different approaches taken to measure and model music similarity in both symbolic and audio domains. Literature relevant to cover song identification can be divided into few areas such as query-by-humming systems, content-based music retrieval, genre classification and audio fingerprinting.

In symbolic domain of cover song identification, symbolic representations of musical content is used in content processing. Query-by-humming systems [4] fall under the symbolic domain as in query-by-humming systems music contents are stored and processed in symbolic representations. This query-by-humming method is parallel to retrieving cover songs from a song database. Even though techniques used in query-by-humming systems could be useful in future approaches of cover song identification, these systems can't achieve high accuracy on real world audio music signals [5], [3].

Audio domain cover song identification approaches focus on measuring similarity of music by exploiting music facets shared between two songs. Extracting invariant features is used to exploit shared music facets. Although such extracted descriptors are responsible to overcome majority of facet changes, special stress is given for achieving tempo, key and structure since those facets are not usually managed by the extracted descriptors themselves [3]. Hence we can look at existing literature in terms of feature extraction, tempo invariance, key invariance, structure invariance and finally similarity comparison. Furthermore, we can take approaches which fall into this general pipeline (refer Figure 4) to look at different techniques used for these stages and distinguish each approach from one another by those techniques.

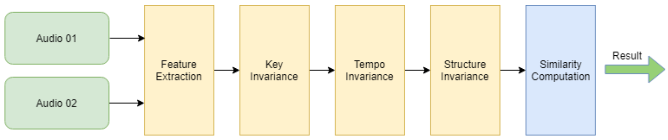

Fig. 4. General Pipeline for Cover Song Identification

Bello's cover song identification method extracts chord sequences as the feature and uses K transpositions for key invariance. Even though there is no technique used for structure invariance, Dynamic Programming (DP) is used for tempo invariance. Finally it uses edit distance to compute the similarity [6]. Since there is no technique used for structure invariance, that method is inefficient against the structural changes in cover songs. Egorov proposed another method which uses the same general pipeline with extracting Pitch Class Profiles (PCP) as the feature. But in this method Egorov uses Optimal Transposition Index (OTI) for key invariance and DP is used for both tempo and structural invariance. And match length is used for similarity computation [7].

Foote [8] and Izmirli [9] introduced two methods which were using Dynamic Time Warping (DTW) for similarity computation and DP for tempo invariance. Both the methods lack techniques for key invariance and structure invariance which makes those methods to perform inefficient in both key and temporal changes. The feature extracted by Foote is energy spectrum while Izmirli extracted key templates. Marlot uses the same techniques for tempo invariance and similarity computation which are DP and DTW, but melody is the extracted feature which uses the key estimation for key invariance.

Related works mentioned above for audio domain can be modelled to the general pipeline for cover song identification (refer Figure 4) as described in Table I.

III. Proposed Method

In the proposed method of remastered song identification, various algorithms are used to extract the audio features, create audio descriptors and match against stored descriptors. Hence we have divided our remastered song identification process in to five steps.

1) Preprocessing
2) Feature Extracting
3) Descriptor Storing (Registering)
4) Matching
5) Postprocessing

Research	Feature	Key Invariance	Tempo Invariance	Structure Invariance	Similarity Computation
Bello [6]	Chords	K transpositions	DP	-	Edit distance
Egorov & Linetsky [7]	PCP	OTI	DP	DP	Match length
Foote [8]	Energy spectral	-	DP	-	DTW
Izmiril [9]	Key templates	-	DP	-	DTW
Marolt [10]	Melody	Key estimation	DP	-	DTW

TABLE I
COVER SONG IDENTIFICATION METHODS AND THEIR TECHNIQUES USED FOR EACH STEP IN GENERAL PIPELINE

Processes of the above steps will be discussed in the following subsections.

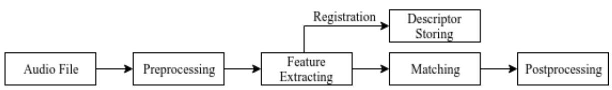

Fig. 5. Remastered Song Identification Process

A. Preprocessing

In default audio data is represented in the time domain. Since even a small change in an audio changes the time domain representation drastically, using the time domain representation of the audio to extract features is not recommended. Hence time domain audio signal is converted to frequency domain signal by using STFT method. STFT is a sequence of Fourier transforms of a windowed signal[11]. 2048 bits long window with 50% overlapping was used as STFT key parameters as depicted in Figure 6.

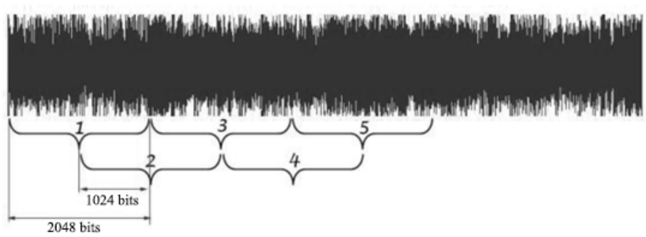

Fig. 6. Key parameters on STFT. 2048 bits long window with 1024 bits long overlapping area.

STFT is often visualized using its spectrogram[11], which is an intensity plot of STFT magnitude over time. The generated spectrogram is converted to a color image as shown in Figure 7. Axis labels and ticks are removed to stop identification of them as key points in feature extracting step.

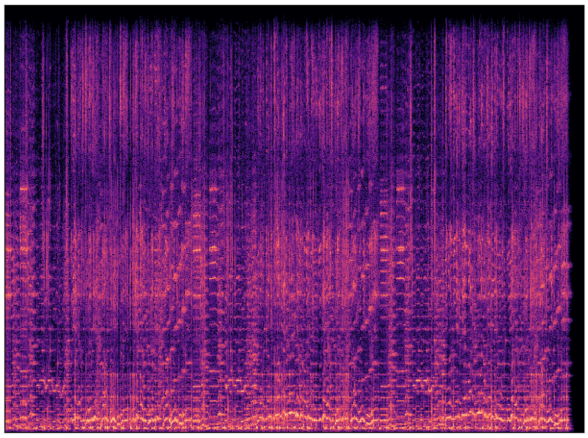

Fig. 7. Generated colour image of spectrogram after preprocessing.

B. Feature Extracting

STFT spectrogram itself can be considered as an audio descriptor[12]. This method uses SIFT[13] to extract the features which are robust to music remastering. In the Figure 8, it can be observed that when tempo is altered the spectrogram will either expand or compress with the time axis and when pitch is altered the spectrogram will either shift upwards or downwards with the frequency axis.

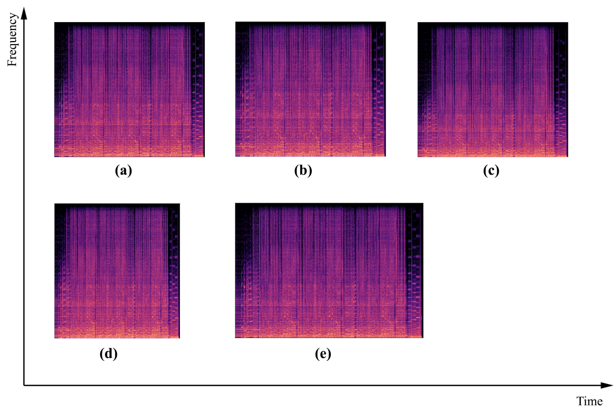

Fig. 8. Spectrogram transformations on audio enhancements. **(a)** is the spectrogram image of a original song. **(b)** 20% pitch increase, **(c)** 20% pitch decrease, **(d)** 20% tempo increase and **(e)** 20% tempo decrease spectrogram images.

SIFT is used in computer vision to identify scale invariant features of an image. SIFT features are invariant to image rotation, scale alterations and illumination[13]. The SIFT feature extractor used in this method consists of four main steps.

1) Scale space extrema detection: Gaussian filters of different scales are applied to the image and potential key points are selected as local minima or maxima of the Difference of Gaussians (DoG) for multiple scales.
2) Keypoint localization: Keypoints that have low contrast or those that are poorly located along edges are filtered out.

3) Orientation assignment: One or more orientations are assigned to each keypoint based on local image gradient.
4) Keypoint descriptor generation: Orientation histograms are created for 4 x 4 pixel neighborhoods for each keypoint. Each histogram consists 8 bins, hence (4 x 4 x 8) 128 dimensional descriptor is generated.

A Set of extracted 128 dimensional descriptors works together in describing the input audio file. Extracted SIFT features are invariant to image stretch and translation which makes them better features to be used in audio identification algorithm which is robust to tempo alterations and pitch shifting.

C. Descriptor Storing (Registering)

SIFT descriptors of original songs must be stored to use them in the matching step of the remastered song identification process. Generally 3-5 minute music clip will have around 2000 key points in its STFT spectrogram. Hence a 2000×128 matrix will be generated for each original song that will be registered.

The descriptor matrix of each original song is converted to a binary string and that binary string is stored in the database as Binary Large Object (BLOB)s. Converting to binary string and storing the matrix as a BLOB will ensure fast recreation of the matrix while retrieving[14].

D. Matching

Music identification is facilitated by matching a feature matrix of a query audio clip with a feature matrix of a original song. Final goal of the matching is to identify the count of keypoints that are matched with the original song. Identification of matching keypoints achieved by taking 2 nearest keypoints to a query keypoint and checking whether the distance to the closest keypoint is lesser than the $0.75 \times$ distance to the 2nd closest keypoint.

E. Postprocessing

the most similar song and matched keypoint count for a given query audio clip is identified in the matching step. But it doesn't exactly mean that query audio clip contains that song. Because the number of keypoints that were matched represents how much the query song matched to the most similar song. Hence there should be a threshold keypoint count to determine whether a query audio clip contains a song in our database or not. But using just a threshold value won't work here since different query audio clips generate different number of key points to match against the database. Hence ratio based threshold is recommended as a measure to determine whether the matched song is actually a correct match. Keypoint ratio can be obtained by the below equation.

$$\text{Keypoint ratio} = \frac{\text{Matched keypoint count}}{\text{Keypoints generated for query audio clip}}$$

Based on this keypoint ratio, a threshold is used to determine the validity of the match found. In order to find this threshold value we have used 844 different audio clips with variable durations to match against 2300 original songs. Those 844 audio clips had 519 audio clips which had songs and 325 audio clips which didn't have songs from those 2300 original songs. And we calculated accuracy for 18 testcases which will be discussed in section IV, and took average of those 18 accuracies for variable threshold values. Then results were illustrated as shown in the Figure 9.

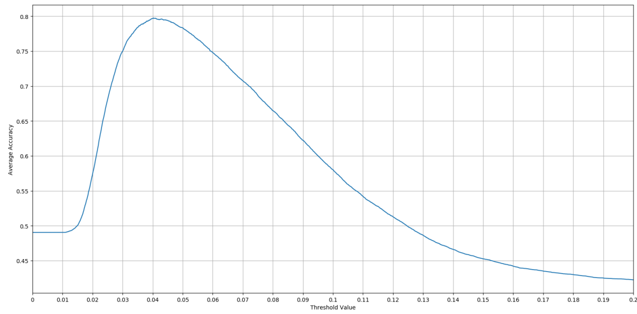

Fig. 9. Average Accuracy Values for Different Threshold Values

Global peak can be observed in the illustration which makes that value a clear threshold point. 0.0298 is the threshold value that was found. Hence if keypoint ratio of a query audio is larger than 0.0298 then it's identified as a valid match to the song that was identified in the matching step, otherwise it's identified as a invalid match. This threshold point makes this method to clearly identify whether a query audio has a song which is in a database or not.

IV. Experiments

A. Song Dataset

Song dataset consisting 2300 sinhala songs were used in the registration step of the experiment. These 2300 sinhala songs were retrieved from Outstanding Song Creators Association (OSCA) of Sri Lanka which works as the governing organization to ensure intellectual property rights of music in Sri Lanka.

B. Query Audio Samples

Variable sized 844 query audio clips were used for the experiment to evaluate the performance of this method against different durations. 519 of the above mentioned audio clips had songs which are in the database while 325 audio clips didn't have songs from the database. Hence for each test case, sample size was 844 query audio clips with 0.61492 prevalence.

C. Test Cases

Test cases were created by doing audio distortions to the query audio samples. Performance of the method was evaluated for three main audio distortions which are tempo alteration, pitch alteration and both pitch and tempo alteration. Both increased and decreased alterations are considered for three different levels of alterations which are 10% alteration, 20% alteration and 50% alteration. Hence there are 3 audio distortions, 2 audio distortion directions and 3 audio alteration levels, 18 ($3 \times 2 \times 3$) test cases were generated to evaluate the performance.

D. *Test Results*

The proposed method has an exact way to identify whether a query audio clip has a matching registered song or not. Therefore this method can be considered as a classifier. A classifier can be evaluated by the confusion matrix generated for a given sample. True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) were calculated for each test case. Then accuracy and FP rate was calculated for each test case. Reducing FP rate is significant as much as increasing the accuracy given that this method mainly focuses on identifying music on radio broadcasts. Accuracy and FP rate can be calculated from the formulas given below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{FP Rate} = \frac{FP}{FP + TN}$$

Keypoint ratio based threshold is proposed as the threshold measure in this method. Two experiments were done, one with keypoint count as threshold measure and the other with keypoint ratio as threshold measure to validate the claim of keypoint ratio is better threshold measure.

1) Using Keypoint Count as Threshold: Average accuracies for different threshold values are illustrated as shown in Figure 10 to identify the optimal threshold value to use. Since the visualization clearly indicates a global peak value which is keypoint count of 69, that value can be used as the threshold value. This threshold value of 69 is used in this experiment.

Results of the experiment using keypoint count as the threshold measure are presented in Table II. There is no clear change between accuracies of pitch changes and tempo changes when using keypoint count as the threshold measure. Pitch changes and tempo changes up to 20% alteration can be identified with 92%-98% accuracy.

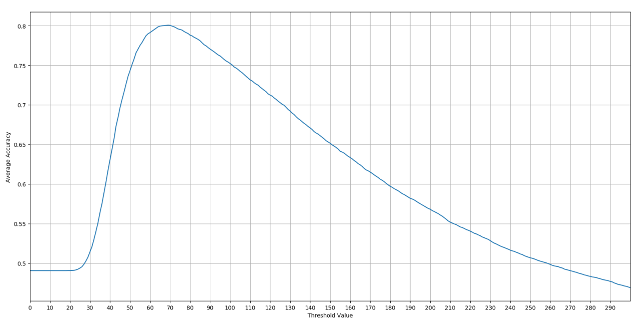

Fig. 10. Average Accuracy Values for Different Threshold Values (Keypoint Count)

Test Case	TP	FP	TN	FN	Accuracy	FP Rate
Tempo Increase 10%	509	19	306	10	0.96564	0.05846
Tempo Increase 20%	509	18	307	10	0.96682	0.05538
Tempo Increase 50%	497	10	315	22	0.96209	0.03077
Tempo Decrease 10%	515	16	309	4	0.97630	0.04923
Tempo Decrease 20%	516	19	306	3	0.97393	0.05846
Tempo Decrease 50%	517	22	303	2	0.97156	0.06769
Pitch Increase 10%	514	21	304	5	0.96919	0.06462
Pitch Increase 20%	513	11	314	6	0.97986	0.03385
Pitch Increase 50%	7	27	298	512	0.36137	0.08308
Pitch Decrease 10%	511	11	314	8	0.97749	0.03385
Pitch Decrease 20%	463	4	321	56	0.92891	0.01231
Pitch Decrease 50%	0	0	325	519	0.38507	0.00000
Tempo & Pitch Increase 10%	414	11	314	105	0.86256	0.03385
Tempo & Pitch Increase 20%	339	18	307	180	0.76540	0.05538
Tempo & Pitch Increase 50%	0	26	299	519	0.35427	0.08000
Tempo & Pitch Decrease 10%	373	5	320	146	0.82109	0.01538
Tempo & Pitch Decrease 20%	226	2	323	293	0.65047	0.00615
Tempo & Pitch Decrease 50%	0	0	325	519	0.38507	0.00000

TABLE II
EXPERIMENT RESULTS USING KEYPOINT COUNT AS THRESHOLD

2) Using Keypoint Ratio as Threshold: Average accuracies for different threshold values are illustrated as shown in Figure 11 to identify the optimal threshold value to use. Since the visualization clearly indicates a global peak value which is keypoint count of 0.0403, that value can be used as the threshold value. This threshold value of 0.0403 is used in this experiment.

Results of the experiment using keypoint ratio as the threshold measure presented in Table III. It can be observed that accuracies have increased considerably compared to results of the experiment which uses keypoint count as the threshold measure which validates the claim of taking keypoint ratio as a better threshold measure. Pitch changes and tempo changes up to 20% alteration can be identified with 95%-99% accuracy.

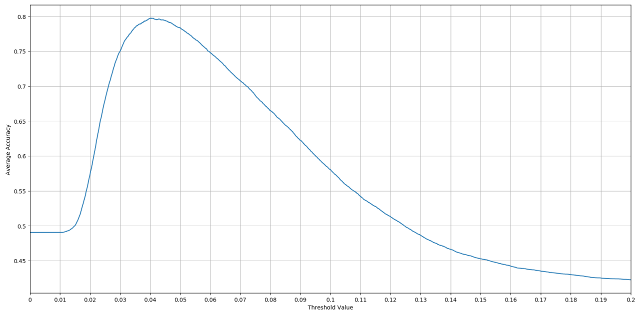

Fig. 11. Average Accuracy Values for Different Threshold Values (Keypoint Ratio)

Test Case	TP	FP	TN	FN	Accuracy	FP Rate
Tempo Increase 10%	512	13	312	7	0.97630	0.04000
Tempo Increase 20%	508	10	315	11	0.97512	0.03077
Tempo Increase 50%	501	9	316	18	0.96801	0.02769
Tempo Decrease 10%	515	11	314	4	0.98223	0.03385
Tempo Decrease 20%	519	10	315	0	0.98815	0.03077
Tempo Decrease 50%	519	9	316	0	0.98934	0.02769
Pitch Increase 10%	519	9	316	0	0.98934	0.02769
Pitch Increase 20%	517	9	316	2	0.98697	0.02769
Pitch Increase 50%	0	4	321	519	0.38033	0.01231
Pitch Decrease 10%	516	16	309	3	0.97749	0.04923
Pitch Decrease 20%	503	25	300	16	0.95142	0.07692
Pitch Decrease 50%	23	160	165	496	0.22275	0.49231
Tempo & Pitch Increase 10%	413	11	314	106	0.86137	0.03385
Tempo & Pitch Increase 20%	287	17	308	232	0.70498	0.05231
Tempo & Pitch Increase 50%	0	7	318	519	0.37678	0.02154
Tempo & Pitch Decrease 10%	432	28	297	87	0.86374	0.08615
Tempo & Pitch Decrease 20%	346	34	291	173	0.75474	0.10462
Tempo & Pitch Decrease 50%	6	154	171	513	0.20972	0.47385

TABLE III
EXPERIMENT RESULTS USING KEYPOINT RATIO AS THRESHOLD

E. Evaluation

Final results of the experiment with keypoint ratio as the threshold measure performs significantly better than the keypoint count threshold measure. Hence, it can be concluded that keypoint ratio is a better threshold measure than keypoint count.

There is a sudden drop of accuracies between 20% and 50% pitch changes. This sudden drop happens when shifting up or down on the frequency axis making spectrogram local patterns to be compressed or expanded which makes the pattern to get unidentified as shown in Figure 12.

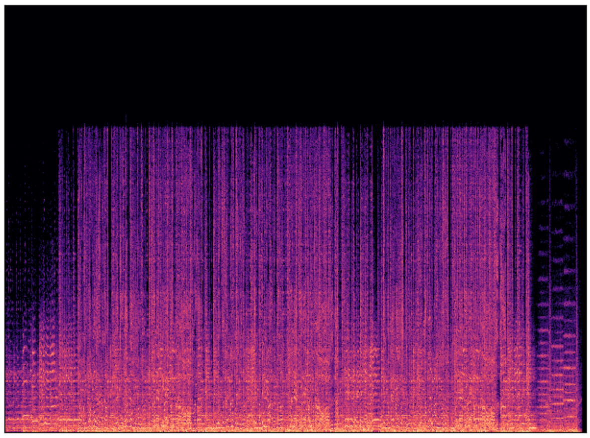

Fig. 12. Generated colour image of a spectrogram (50% pitch decrease)

V. Conclusion

In this paper, a novel music similarity descriptor is proposed. Using SIFT descriptors to match two STFT spectrograms suits identification of remastered songs as it can be used to match variable sized audio clips with original songs. By using the new descriptor above 96% accuracy can be achieved for tempo alterations up to 50% and above 95% accuracy can be achieved for pitch alterations up to 20%. Which makes it robust to audio enhancements performed on music.

References

[1] Parliament of the democratic socialist republic of Sri Lanka, "Intellectual Property Act, No.36 of 2003."

[2] E.D.N.W. Senevirathna and K.L. Jayaratne, "Radio Broadcast Monitoring to Ensure Copyright Ownership," *International Journal on Advances in ICT for Emerging Regions*, vol. 11, 2018.

[3] J. Serrà, E. Gómez, and P. Herrera, "Audio cover song identification and similarity: Background, approaches, evaluation, and beyond," in *Studies in Computational Intelligence*, 2010, vol. 274, no. 1, pp. 307–332.

[4] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database," in *Proceedings of the third ACM international conference on Multimedia - MULTIMEDIA '95*. San Francisco, California, United States: ACM Press, 1995, pp. 231–236. [Online]. Available: http://portal.acm.org/citation.cfm?doid=217279.215273

[5] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the MUSART testbed," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 687–701, Mar. 2007. [Online]. Available: http://doi.wiley.com/10.1002/asi.20532

[6] J. P. Bello, "Audio-Based Cover Song Retrieval Using Approximate Chord Sequences: Testing Shifts, Gaps, Swaps and Beats." in *ISMIR*, vol. 7. Citeseer, 2007, pp. 239–244.

[7] A. Egorov and G. Linetsky, "Cover song identification with IF-F0 pitch class profiles," *MIREX extended abstract*, 2008.

[8] J. Foote, "ARTHUR: Retrieving Orchestral Music by Long-Term Structure." in *ISMIR*, 2000.

[9] Özgür İzmirli, "Tonal Similarity from Audio Using a Template Based Attractor Model." in *ISMIR*. Citeseer, 2005, pp. 540–545.

[10] M. Marolt, "A Mid-level Melody-based Representation for Calculating Audio Similarity." in *ISMIR*, 2006, pp. 280–285.

[11] N. Kehtarnavaz, "Frequency Domain Processing," *Digital Signal Processing System Design*, vol. 1, pp. 175–196, 2008.

[12] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. I, pp. 597–604, 2005.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] R. Sears, C. Van Ingen, and J. Gray, "To BLOB or Not To BLOB: Large Object Storage in a Database or a Filesystem?" Tech. Rep., 2006.