Representative Music Fragments Extraction by Using Segmentation Techniques

Wen-Jie Ke, Chuan-Wang Chang, and Hewijin Christine Jiau

Department of Electrical Engineering
National Cheng Kung University
chuan@ee.ncku.edu.tw

Abstract

Melody extraction is an important issue of research for successful development of music information retrieval system, particularly for polyphonic music. The extracted melody can further processed as music index that can speed up the retrieval of large music collections. In the past, most researchers assume music is monophonic. If the music is polyphonic, there is not satisfyingly solved by existing known algorithms. In this paper, we propose two methods: "Four-Measure Based segmentation" and "LBDM Concatenation" to extract the representative fragments of melody from different tracks. We also analyze the effects caused by the varied combinations of music features to help us extract the melody.

Keywords: melody extraction, polyphonic music, LBDM, music information retrieval.

1 Introduction

With the development of multimedia technology, content-based music retrieval has attracted much interest in recent years. It allows users query by music content rather than metadata. To achieve this task, techniques for matching melody fragments are required. Besides, a good melody extraction approach affects the quality of music retrieval seriously. The ability to identify and track the melody line of a musical piece could be very useful for music retrieval. Melody extraction also can be applied to the representation of a musical piece: like a summary for a text, the melody line or significant parts therefore can be seen as an index of a musical piece.

Most of the human music listeners, even without musical education, are able to track the melody line of a selected musical instrument and the main melody of a polyphonic musical piece - even without knowing which type of instrument or which voice will represent the main melody. However, it seems also to be the most challenging part of retrieval and only very few researches has presented in this area. This is because of the difficulty to extract melodic information from polyphonic music.

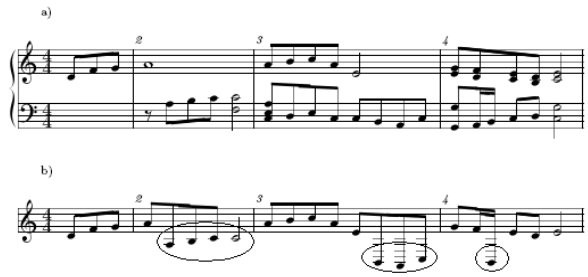

Figure 1. (a) A short piece of music. (b) The extracted monophonic melody by combining all tracks and keeping the highest note from all simultaneous note events.

In the past, researchers [1] [2] [3] extract melody form music manually or semiautomatically for further processing. They assume the music is monophonic, because their focus is on indexing technique or similarity matching. However, the majority of music is polyphonic. When the number of song increases explosively, it is hard to extract melody of music manually. Hence, methods for automatic melody extraction are required. Ghias [4] thought of that, but he just simply discards percussion. Uitenbogerd [5] [6] provided a method to extract melody from polyphonic music. He combines all tracks and keeps the highest note from all simultaneous note events. Shan [7] adds the feature of volume in his analysis. The reason is that volume of melody is usually the largest. There have problems for these two methods: they will contain extra and incorrect notes.

That is, the extracted melody may contain the accompaniment notes. In Figure 1, the marked circles show the redundant notes of the extracted melody. Therefore, Chen *et al.* [8] only select one track as melody. They use *pitch density* as the feature for melody selection. Tang [9] use *AvgVel, PMRatio, SilenceRatio, Range,* and *TrackName* to select one track as melody from polyphonic music. They often assume that the melody is only on one track. In reality, the melody may move from one track to another.

The rest of this paper is organized as follows. In the following section, we will introduce the proposed methods. Then in Section 3, we describe the experiments and we discuss the accomplished experiments in Section 4. We summarize the paper and outline our future work in Section 5.

2 Proposed Method

As discuss above, we see the drawbacks of existing works: they assume melody is only on one track. In reality, melody usually separated into many fragments that are located on different tracks. As shown in Figure 2, the music excerpt composes of three melody fragments that distributed over three tracks. The number in marked part indicates the ratio of melody fragment to whole melody. By using track as the unit of analysis, we can get portion of the whole melody, i.e., we will lose many significant fragments.

Track1	0.375				
Track2			0.375		
Track3				0.25	
Track4					

Figure 2. An example of melody fragments that distributed over three tracks.

In order to improve such a condition, we use a smaller unit instead of track. In musicology, *phrase* is an important structure of music, which expresses a complete thought of music. It can be treated as an independent fragment of music [10]. Figure 3 shows an example of phrases. The phrase usually ends with rest note or note with long duration. For this reason, choosing phrase as the unit of analysis is preferred.

2.1 Four-measure Based Segmentation

In musicology, a "phrase" is usually a unit with four measures [11]. In most parts of folk songs, it usually uses four measures as a phrase. Therefore, the size of four measures is selected as the basic unit for further

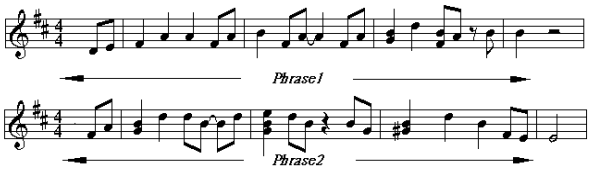

Figure 3. A music excerpt with two phrases.

experiments. In this approach, music is partitioned into several fragments with the unit of four measures.

But not all of the songs are using four measures as a phrase. It may be six or eight measures, even an unfixed one. Hence, another approach for selecting the changeable phrase is required.

2.2 LBDM Concatenation

The Local Boundary Detection Model (LBDM) [12] [13] enables the detection of local boundaries in a melodic surface and can be used for musical segmentation. By using LBDM, we can get the changeable phrases. The LBDM uses duration, pitch, and rest note to separate the music into smaller pieces. However, the LBDM fragments sometimes lack of perceptual accuracy. Here, we propose an improved method that considers the duration and rest note of LBDM fragments to get a more complete phrase. This method is called *LBDM Concatenation*.

The main strategy of the LBDM Concatenation method consists of the following three steps:

- Step 1. If the ending note of the LBDM fragment is a long rest note and the duration is equal or greater than the half of duration of one measure, it should be regarded as the end of the phrase. So, this fragment will be concatenated with the preceding one. However, if the preceding fragment's ending note is also a long rest note (equal or greater than the half of duration of one measure), don't concatenate it.

- Step 2. For the remaining LBDM fragments, if the last note's duration of fragment (not include the rest node) is equal or smaller than the last note's duration of next fragment, these two fragments may be concatenated.

- Step 3. After the previous steps, a lot of more complete phrases are shown, but some concatenated fragments are still too small, they have about one or two measures. Hence, in the last step, concatenating the fragments whose duration is not

Definition:
P_i: pitch of last note of ith LBDM fragment
D_i: duration of last note of ith LBDM fragment
D_s: duration of one measure of this song
N_i: number of LBDM fragments before step i
TD_i: total duration of ith LBDM fragment

Input: LBDM fragments
Output: Fragments with more complete phrase

Begin
  //Step 1: using rest note
  for $i = 1$ to N_1
    if P_{i+1}=0 and $D_{i+1} \geq D_s/2$ and $D_i \leq D_s/2$
      Concatenate D_i and D_{i+1}
    end if
  end

  //Step 2: using duration
  for $i = 1$ to N_2
    if $D_i \leq D_{i+1}$ and $D_i \leq D_s/2$
      Concatenate D_i and D_{i+1}
    end if
  end

  //Step 3: let the duration of
  //fragment over four measures
  for $i = 1$ to N_3
    if $TD_i < 4D_s$ and $D_i < D_s$
      Concatenate D_i and D_{i+1}
    end if
  end
End

Figure 4. The high-level description of the LBDM Concatenation algorithm.

exceeds four measures. If the ending note's duration is greater than the fully rest note, do not concatenate it. Because this fragment is complete.

The LBDM Concatenation algorithm is shown in Figure 4. We explain the algorithm by giving an example. Consider the original music excerpt as shown in Figure 5. It partitioned into twelve LBDM fragments. The numbers over the track are serial number of LBDM fragments.

In Step 1, because of the ending note of the fourth LBDM fragment is a long rest note and the duration is equal to a half of one measure, we regard fragment 4 as the end of a phrase. Hence, we concatenate fragment 4 with fragment 3, as shown in Figure 6.

In Step 2, since the last note's duration of fragment 2 (not includes the rest node) is smaller than the last note's duration of fragment 3, we concatenate these two fragments. We deal with fragments 6, 7, 8, 10, and 11 by the same manner, as shown in Figure 7.

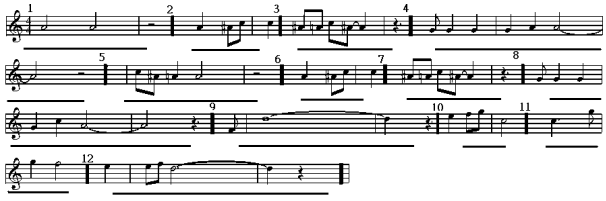

Figure 5. The original LBDM fragments.

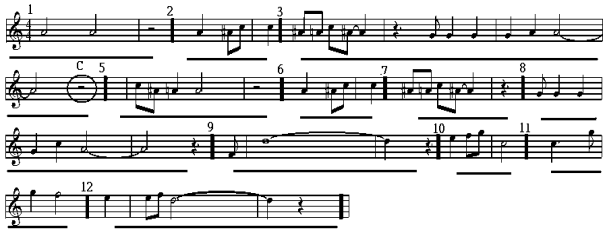

Figure 6. In Step1, concatenate fragment 4 with fragment 3.

In Step 3, to get a longer phrase, we concatenate fragment 1 with fragment 2. Fragment 5 and fragment 6 are also concatenated. Figure 8 shows the final phrases. After processing, it only remains three fragments, and these concatenated fragments are perceptually similar to the complete phrases.

3 Experiment Set-Up

A collection of 80 MIDI files of Taiwanese popular song obtained from Web is used for verifying the three approaches: Track based, Four-Measure based, and LBDM Concatenation based. For the experiments, 40 songs are random selected as the learning data and 40 songs as the testing data. We separate 40 testing songs into two groups: "Single-track" and "Multi-track". Each group has 20 songs. Single-track means melody is only on one track, while Multi-track means the melody is distributed over different tracks. The reason of separating the songs into Single-track and Multi-track is to verify the prior work is not good enough for melody selection on Multi-track. Here are the steps of the experiments.

- Step 1. To obtain the difference between representative part and non-representative part, we extract the representative part of 80 songs manually.

- Step 2. In order to get more precision phrase, we investigate several features of music. In our experiments, we consider nine features, like [14], for the analysis. Furthermore, we verify the influence

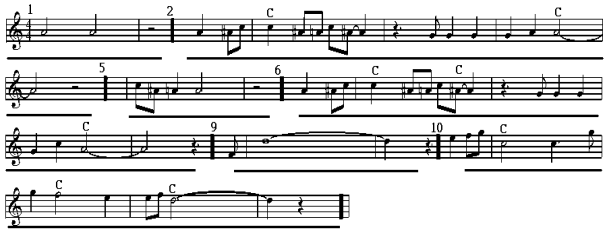

Figure 7. In Step2, fragments 6, 7, 8, 10, and 11 are processed.

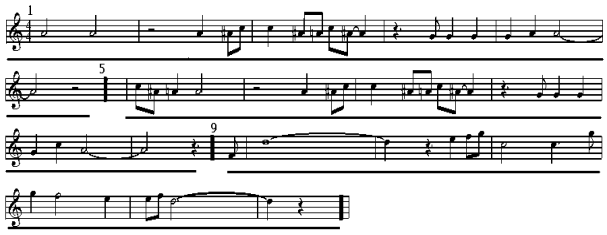

Figure 8. The last phrases obtained by LBDM Concatenation.

of these nine features. These features are *Kinds of Pitches*, *Average Pitch*, *Pitch Variance*, *The difference between highest pitch and lowest pitch*, *Average of total difference between pitches*, *Average number of Notes*, *Kinds of Duration*, *Duration Variance*, and *Time Distribution (Total duration of notes / Track Duration)*.

- Step 3. In learning phase, we estimate the average of each feature for representative part and non-representative part respectively. Because each feature does not have the same importance for selecting the melody of music, the weight of each feature is also decided. Figure 9 is the distribution of "Average Pitch" for representative part and non-representative part. The weight is defined as

$$W_i = \frac{A_i}{TA_i} \tag{1}$$

where A_i indicates area of representative part above non-representative part, and TA_i denotes total area of representative part. If both representative and non-representative parts are totally disjointed, then the importance of this feature is 100%. In this case, the weight of Average Pitch is 0.731.

- Step 4. For testing songs, we introduce the Weighted Euclidean Distance to estimate the distance between testing data and learning result of

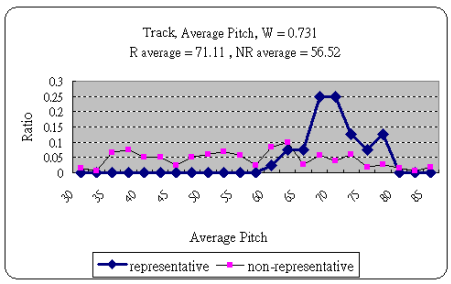

Figure 9. Distribution of average pitch.

representative part. The Weighted Euclidean distance is defined as

$$d_i = \sqrt{\sum_{i=1}^{n} W_i (F_i - A_i)^2} \tag{2}$$

where W_i indicates $feature_i$'s weight of representative part, F_i denotes feature's value of testing part, and A_i is the average of $feature_i$ of representative part. If the distance is close to the learning result of representative part, the testing data is similar to the representative part.

- Step 5. Comparing the distances of all the tracks and selecting the minimum as the representative fragment. Because of the minimum distance means the most similarity between the testing data and representative part. For example, Figure 10 shows the Weighted Euclidean Distance of each track. The Track 3 has the lowest value of all tracks. We select track 3 as the representative part of this song.

- Step 6. In this step, we estimate the accuracy. If the selected representative fragment is coincides with fragment that selected by human, we call this selected one is *correctness*. We regard the ratio of the correctness count to the total representative fragments as the accuracy. It is defined as

$$Accuracy = \frac{CR}{TR} \tag{3}$$

Track#	Weighted Euclidean Distance
Track1	1.017
Track2	1.578
Track3	0.578
Track4	0.876
Track5	0.958

Figure 10. The Weighted Euclidean Distance of each track.

Table 1. The Learned Average of Nine Features and Their Weights

Track	Pkinds	Dkinds	PitchDiff	AveNotes	AvePitch	Pvariance	Dvariance	AvePitchDiff	Time
Representative	13.575	16	22.575	4.814	71.102	24.417	16.775	2.871	0.851
Non-Representative	8.957	8.492	15.403	2.877	56.52	19.028	22.704	2.848	0.563
Weight	0.561	0.612	0.576	0.715	0.731	0.611	0.667	0.657	0.586
Four Measures	**Pkinds**	**Dkinds**	**PitchDiff**	**AveNotes**	**AvePitch**	**Pvariance**	**Dvariance**	**AvePitchDiff**	
Representative	5.944	4.596	11.22	5.104	70.56	13.26	17.01	2.56	
Non-Representative	4.42	3.38	8.11	5.327	53.34	11.17	21.09	2.53	
Weight	0.467	0.419	0.351	0.265	0.627	0.298	0.331	0.379	
LBDM	**Pkinds**	**Dkinds**	**PitchDiff**	**AveNotes**	**AvePitch**	**Pvariance**	**Dvariance**	**AvePitchDiff**	
Representative	6.51	4.907	11.02	5.267	69.35	12.95	16.02	2.55	
Non-Representative	4.88	3.99	9.07	6.953	55.51	12.9	20.75	2.53	
Weight	0.392	0.248	0.32	0.2496	0.525	0.313	0.41	0.415	

where CR means the correctness count of representative fragments, and TR means the total representative fragments. For example, there are 20 songs with melody only on one track. If it gets 14 correct melody lines, the accuracy is 0.7.

- Step 7. For all the features, there have no idea whether it is good enough to put them together. Therefore, we conduct many experiments that consider varied combinations of these nine features.

4 Experimental Results and Observations

Table 1 shows the learned value of each feature by using three methods. The values of representative part are higher than non-representative part, except Dvariance. In music, representative part often gets more changeable in pitch and duration than nonrepresentative part. The experimental results coincide with the phenomenon. However, what happened for Dvariance? We observe that most non-representative tracks have fewer notes, and the duration of note sometimes has greater change. Although representative parts also have greater change, they have more notes than non-representative parts. We can understand this by observe the average of AveNotes. Therefore, because of owning more notes, the value of Dvariance of representative part is lower than non-representative part.

When the size of unit reduced from track to four measures, the value of each feature become smaller, except AveNotes. It is because the non-representative part usually contains many rest notes. The rest notes will be ignored in analysis. Hence, the value of AveNotes of non-representative part is higher than representative part.

For the LBDM Concatenation method, duration of fragment is usually more than four measures. Because of the unit of analysis is longer than Four-Measure, so for Dkinds, Pkinds, and AveNotes, the value of representative part is greater than non-representative part.

Table 2 and Table 3 show the combinations of varied features by using Four-Measure based method and LBDM Concatenation method respectively. In the case of Four-Measure based, when removing the effect of Pvariance, it will get better result than consider all nine features. In the case of LBDM Concatenation method, if we remove the Dvariance from other features, we get a higher accuracy in S-Track while get a lower accuracy in M-Track.

Table 2. Varied Combination of Features for Four-Measure

Pkinds	Dkinds	PitchDiff	AveNotes	AvePitch	Pvariance	Dvariance	AvePitchDiff	S-Tracks	M-Tracks	Average
*	*	*	*	*	*	*	*	0.6083086	0.4115854	0.509947
	*	*	*	*	*	*	*	0.6053413	0.4115854	0.50846335
*		*	*	*	*	*	*	0.6053413	0.4115854	0.50846335
*	*		*	*	*	*	*	0.6053413	0.4085366	0.50693895
*	*	*		*	*	*	*	0.6053413	0.402439	0.50389015
*	*	*	*		*	*	*	0.6023739	0.4085366	0.50545525
*	*	*	*	*		*	*	0.7507418	0.4847561	0.61774895
*	*	*	*	*	*		*	0.6212351	0.4115854	0.51641025
*	*	*	*	*	*	*		0.6023739	0.4085366	0.50545525
*		*	*	*	*		*	0.7418398	0.4847561	0.61329795
*	*		*	*			*	0.7002968	0.4512195	0.57575815
*	*	*	*	*			*	0.7121662	0.4664634	0.5893148
*			*	*	*		*	0.7240356	0.472561	0.5982983

Table 3. Varied Combination of Features for LBDM Concatenation

Pkinds	Dkinds	PitchDiff	AveNotes	AvePitch	Pvariance	Dvariance	AvePitchDiff	S-Tracks	M-Tracks	Average
*	*	*	*	*	*	*	*	0.7408152	0.549982	0.6453986
	*	*	*	*	*	*	*	0.7408152	0.5394023	0.64010875
*		*	*	*	*	*	*	0.7370109	0.5382791	0.637645
*	*		*	*	*	*	*	0.7441486	0.5438407	0.64399465
*	*	*		*	*	*	*	0.7474819	0.5482066	0.64784425
*	*	*	*		*	*	*	0.7372463	0.4487681	0.5930072
*	*	*	*	*		*	*	0.7512354	0.544982	0.6481087
*	*	*	*	*	*		*	0.7683515	0.5336233	0.6509874
*	*	*	*	*	*	*		0.7408152	0.5374639	0.63913955
	*	*	*	*	*		*	0.776558	0.5336233	0.65509065
*	*		*	*	*		*	0.7816123	0.5213769	0.6514946
*	*	*	*	*	*			0.771395	0.5336233	0.65250915

Figure 11 summarizes the accuracy of three methods. If the testing data with the melody in one track, experimental results show that the proposed LBDM Concatenation method correctly identifies the melody in 78% of the test cases. The Four-Measure Based method correctly identifies the melody in 75%. If the testing data with melody distributed over multi-track, the accuracy of LBDM concatenation method is 53%, better than the accuracy 8% of prior work. Both our proposed methods get a higher accuracy than prior work such as Track based method. This is because, by using track as the unit of analysis, we can get portion of the whole melody, i.e., we will lose many significant fragments.

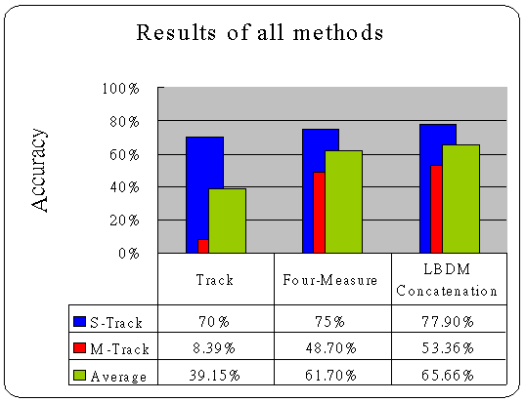

Figure 11. Comparison of all methods.

5 Conclusion and Future work

Our analysis is useful in advanced application domains such as music information retrieval. In this paper, we point out the shortcomings of the existing works and verify issues by experiments. We propose two methods for melody extraction. They are "Four-Measure based segmentation" and "LBDM Concatenation", respectively. Because the size of unit for analysis is reduced, more representative fragments for which melody is separated into Multi-track are obtained. Experiments show that the results are better than prior works.

Experimental results show that if using this LBDM Concatenation method, we can extract a more precise "phrase". Furthermore, the extracted representative fragments can further process as index of song. As the result, the searching time and required storage can be dramatically decreased.

The ultimate application of this work would be to implement an interactive system for music information retrieval.

References

[1] J. L. Hsu, C. C. Liu, and A. L. Chen, "Discovering Nontrivial Repeating Patterns in Music Data," *IEEE Transactions on Multimedia*, vol. 3, no. 3, pp. 311–325, 2001.

[2] C. W. Chang and H. C. Jiau, "Extracting Significant Repeating Figures in Classic Music by Using Quantized Melody Contour," *Proceedings of IEEE International Symposium on Computers and Communications*, pp. 1061–1066, June 2003.

[3] C. W. Chang and H. C. Jiau, "An Improved Music Representation Method by Using Harmonic-Based Chord Decision Algorithm," *Proceedings of IEEE International Conference on Multimedia and Expo*, June 2004.

[4] A. Ghias, J. Logan, and D. Chamberlin, "Qurey by Humming - Musical Information Retrieval in an Audio Database," *Proceedings of the 3rd ACM International Conference on Multimedia*, pp. 231–236, 1995.

[5] A. Uitdenbogerd and J. Zobel, "Manipulation of Music for Melody Matching," *Proceedings of ACM International Multimedia Conference*, pp. 235–240, September 1998.

[6] A. Uitdenbogerd and J. Zobel, "Melodic Matching Techniques for Large Music Databases," *Proceedings of ACM International Multimedia Conference*, pp. 57–66, November 1999.

[7] M. K. Shan, F. F. Kuo, and M. F. Chen, "Music Style Mining and Classification by Melody," *IEICE Transactions on Information and Systems*, vol. 1, pp. 26–29, August 2002.

[8] H. C. Chen and A. L. Chen, "A Music Recommendation System Based on Music Data Grouping and User Interests," *Proceedings of the 10th International Conference on Information and Knowledge Management*, pp. 231–238, 2001.

[9] M. Tang, C. L. Yip, and B. Kao, "Selection of Melody Lines for Music Databases," *The 24th Annual International Computer Software and Applications Conference*, pp. 243–248, October 2000.

[10] Z. Q. Wu, *A Guide to Musical Forms*. Mercury Publishing House, 1996.

[11] L. Stein, *Structure and Style–The Study and Analysis of Musical Forms*. Evanston, Ill., Summy-Birchard Company, 1962.

[12] E. Cambouropoulos, "A Formal Theory for the Discovery of Local Boundaries in a Melodic Surface," *Proceedings of the III Journees d' Informatique Musicale*, 1996.

[13] E. Cambouropoulos, "The Local Boundary Detection Model (LBDM) and its Application in the Study of Expressive Timing," *Proceedings of the International Computer Music Conference*, pp. 17–22, September 2001.

[14] E. G. Gutierrez, *Melodic Description of Audio Signals for Music Content Processing*. A thesis submitted for the degree of Doctor of Philosophy, 2002.