Development of Audio Melody Extraction and Matching Engine for MIREX 2011 tasks

송재종 장달원 이석필 *박호종
전자부품연구원 정보통신미디어본부 디지털미디어연구센터
*광운대학교 전자공학과
jcsong@keti.re.kr

Development of Audio Melody Extraction and Matching Engine for MIREX 2011 tasks

Song, Chai-Jong Jang, Dalwon Lee, Seok-Pil *Park, Hochong
DigitalMedia R&D Center, Broadcasting and ICT R&D Division, KETI
*Dept. of Electronics Engineering, Kwangwoon University

Abstract

In this paper, we proposed a method for extracting predominant melody of polyphonic music based on harmonic structure. Harmonic structure is an important feature parameter of monophonic signal that has spectral peaks at the integer multiples of its fundamental frequency. We extract all fundamental frequency candidates contained in the polyphonic signal by verifying the required condition of harmonic structure. Then, we combine those harmonic peaks corresponding to each extracted fundamental frequency and assign a rank to each after calculating its harmonic average energy. We run pitch tracking based on the rank of extracted fundamental frequency and continuity of fundamental frequency, and determine the predominant melody. For the query by singing/humming (QbSH) task, we proposed Dynamic Time Warping (DTW) based matching engine. Our system reduces false alarm by combining the distances of multiple DTW processes. To improve the performance, we introduced the asymmetric sense, pitch level compensation, and distance intransitiveness to DTW algorithm.

1. Introduction

In order to analyze and search musical data efficiently, there has been growing interest in computational auditory scene analysis (CASA), multi-pitch extraction, QbSH and music information retrieval (MIR). To improve MIR technologies, music information retrieval evaluation exchange (MIREX) is suggested by Stephen Downie in 2004 and started from 2005 as named audio description contest (ADC). They changed the contest name to MIREX in 2006. There are various tasks related with MIR. Some tasks are vanished or joined by agreement of community annually. We are interested in QbSH and audio melody extraction (AME) task among them [4-8]. So we proposed and submitted algorithms for QbSH and AME task in MIREX 2011.

To process of extracting the predominant melody or vocal melody from polyphonic music is required commonly for those methods, and a great deal of research is being carried out to this end. The previous technologies can be classified into the decomposition method for spectral parameters and the modeling method using the statistical properties[1][2]. Spectral decomposition method extracts the melody without probability modeling of signals by using spectrum harmonics, frequency, timbre characteristics, non-negative matrix factorization (NMF) and filter bank. The existing technologies so far have differences in that they implement the same goal in different ways. The proposed method proposed in this paper can search the more accurate fundamental frequency since it obtains all the fundamental frequency (F0) using all given peak positions. In addition, the proposed method is a simple and unsupervised method, and shows high accuracy using the harmonic structure of spectrum without complex processing of signal modeling.

2. Proposed Melody extraction method

2.1 F0 Candidates extraction

Since polyphonic music contains multiple sound sources at the same time, in order to extract the main melody, it is necessary to extract multi-pitch frequency first, and proceed to the process of selecting F0 corresponding to the main melody out of the extracted multi-pitch frequency. In order to implement these actions effectively, in this paper, we propose multi-pitch extraction and main melody extraction method based on the harmonic structure that is important characteristic of musical signals.

Since the fundamental frequency of musical signal is determined by the low frequency band, the high-bandwidth component does not affect our extracting F0. Therefore, input music signal is down sampled to 8 kHz sampling frequency in the preprocessing module.

Multi-pitch extraction module extracts several pitch

information included in the signal, select the valid pitch according to the availability and accuracy of harmonic structure of the extracted pitch and then determine the rank of each pitch. First, this method search all possible meaningful candidates for F0 contained in the signal. By searching the frequency peaks with high probability of harmonic peak, and analyzing the gap between the peaks, it determines whether the relevant peaks meet the conditions of harmonic structure and finds F0 satisfying the conditions. Then, Harmonic structure clustering module combines each harmonic component for all F0 candidates and generates the harmonic group. Finally, it determines the priority ranking of each F0 by calculating the average energy of each harmonic group.

Pitch tracking module runs pitch tracking by considering the fundamental frequency continuity of the frames before and after and considering the rank of fundamental frequency, and it selects the final fundamental frequency corresponding to the main melody.

2.2 Pitch Tracking

We select predominant F0 through pitch tracking of F0 candidates extracted by each frame. Pitch tracking module runs based on the continuity of F0 between frames around and its processing is as follows:

- Measure the F0 continuity with the previous frame and the next frame based on the first order F0 of current frame.
- If the first order F0 of current frame is not continuous and F0s of the previous frame and the next frame are identical, then take F0 of current frame as F0 of the previous frame.
- If the first order F0 of current frame is not continuous and F0s of the previous frame and the next frame are not identical, then take the continuous F0 as F0 of the previous frame by measuring the continuity of 2nd and 3rd order F0s of current frame.
- If it is not processed above, determine the first order F0 of current frame as starting point of a new sound.

2.3 Post-processing

Pitch doubling and halving errors are a common problem in pitch extraction, and pitch errors are suddenly raised. Post-processing is implemented based on vowels and instrument sounds are stable signal, the melody line doesn't depart from criteria scale. Its processing should be done as follows:

- Using the conditions of equation (1), we cluster F0, where G_l is high similarity F0 of group, l is group index, n is frame index, f_n is threshold value of group. β is setting on at 1.5 tone by considering vibration. We determine high trusted group by group length, and F0 of this group is criteria F0. If around satisfy at integer multiple of the criteria F0, F0 of short term group converge criteria F0.
- We set Median F0 is criteria F0, and Using the conditions of equation (1), we cluster F0. β is setting on at 1 tone. If the difference between F0 of all group and criteria F0 is more than 1.5 octave, F0 of group is completely wrong.

$$\begin{array}{ll} \mathrm{if}(\beta > |f_n - f_{n+1}|) & f_n \in G_l \\ else & l++ \end{array} \quad (1)$$

2.4 Voiced frame detection

After post-processing, we work Voiced frame detection processing. We use the conditions of equation (1), cluster F0, and recalculate harmonic average energy. Also, we use harmonic sharp of cluster. Then we use harmonic average energy and harmonic sharp which are suitable to distinguish vocal and non-vocal frames.

3. Proposed Matching Engine

Matching engine of our QbSH system is based on DTW algorithm, and we incorporate the combination of DTW distance, asymmetric sense, compensation, and distances insensitive to the error. To reduce false alarm, various DTW processes are performed, and the DTW distances are combined. To match a short query with a song, asymmetric DTW is used. The method to compensate the incorrect singing/humming and the saturated distances which are not highly sensitive to the error, are also used in the system.

The input of QbSH matching engine is a set of frame-based pitch sequences. Matching engine should be robust against the mismatch between pitch sequence of query Sq and the pitch sequence stored in DB. From now, $S_{DB}^{(i)}$ denotes the pitch sequence of ith song. The objective of matching engine can be mathematically formulated as follow:

$$\hat{i} = \arg\min_{i=1}^{I} d_M(S_q, S_{DB}^{(i)}) \quad (2)$$

where $d_M(*)$ is the distance computed in matching engine.

Commonly, users sing/hum at inaccurate absolute/relative pitch with a wrong tempo [6]. To compensate pitch, brute-forth search is used. The system finds the minimum distance by changing the compensation coefficient. Mathematically,

$$\begin{aligned} d_M(S_q, S_{DB}^{(i)}) = {} & \gamma_1 \min_{c \in C_1} d_{DTW}(S_q + c, S_{DB}^{(i)}) + \\ & \gamma_2 \min_{c \in C_2} d_{DTW}(S_q + c, S_{DB}^{(i)}) + \\ & \gamma_3 d_{DTW}(S_q, S_{DB}^{(i)}) \end{aligned} \quad (3)$$

where c is compensation coefficient and $(\gamma_1, \gamma_2, \gamma_3)$ are weighting coefficients. The sequence $\widehat{S_q}$ and $\widehat{S_{DB}^{(i)}}$ mean delta sequence of S_q and $S_{DB}^{(i)}$, respectively. In our system, C_1 is set to $\min(S_{DB}^{(i)}) - \min(S_q) + (-5, -4, ..., 5)$, and C_2 is set to $\min(S_{DB}^{(i)}) - \min(S_q) + (-4.5, -3.5, ..., 4.5)$. Thus, $d_M(*)$ is the combination of three distances. Among the three, the first two are decided based on minimum search. By combining the distances, the false alarm is reduced. Our system uses the DTW algorithm which is widely used for QbSH system since it gives the robust matching results against local timimg variation and inaccurate tempo.

$$d_{HINGE}^{(\lambda)}(a,b)=\begin{cases}|a-b| \text{ if } |a-b|<\lambda \\ \lambda \quad otherwise\end{cases} \tag{4}$$

As in [7], determining DTW path is asymmetric, and the difference is a weighting coefficient. In system proposed in [7], the weighting coefficient is 2, but we set it as 4. The performance of QbSH system is dependent on the distance between two elements of different vectors. When commonly used. In our works, the following distance is used.

For the second subtask, we should keep it in mind that two pitch sequences from singing/humming snippets may be very different for a short period. To reduce the influence of the short period error, the distance is used. In the distance, very big difference is limited as only λ. It is set to 3 for our system.

4. Conclusion

MIREX evaluated all of algorithms and systems which are submitted for every tasks with various dataset and announced the results. There are 4 different dataset for AME task. The performance is measured by overall accuracy (OA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), and voiceing false alarm (VFA) with each dataset [3]. RCA and OA represents the accuracy of measured melody frequency for the groundtruth and is define by equation (5).

$$RPA=\frac{TPC+FNC}{GV}$$
$$OA=\frac{TPC+TN}{TO} \tag{5}$$

YSLP1 which is name of algorithm we proposed for AME task won the ADC 2004 dataset in MIREX 2011.

Table 1. Result for AME task with ADC 2004 dataset

Algorithm	OA	RPA	RCA	VFA
YSLP1	85.33	86.72	86.96	29.73
PJY1	80.69	84.88	87.59	29.87
SG2	73.97	77.28	79.41	15.25
SG1	73.55	76.34	78.71	15.09
LYRS1	73.03	84.51	86.16	87.37
CWJ1	72.73	73.08	76.64	29.28
TOS1	59.42	73.03	81.43	29.37
TY3	46.99	56.4	65.33	93.69
TY4	46.99	56.39	65.55	93.69
HCCPH1	44.13	41.69	53.95	27.17

For the QbSH task, MRR is measured by equation (6).

$$MRR=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{rank_n} \tag{6}$$

where, N is number of songs within DB and $rank_n$ is the ranking score of returned list with queried humming/singing.

Table 2. Result for AME task with various dataset

Audio Melody Extraction				
		Overall Accuracy		
SubID	Participants	MIREX'09 0db	MIREX'09 +5db	MIREX'09 -5db
SG1	Salamon, Gómez	**0.78**	0.85	0.61
SG2	Salamon, Gómez	0.78	**0.85**	0.61
TOS1	Tachibana, Ono, Ono, Sagayama	0.74	0.82	**0.62**
PJY1	Park, Jo, Yoo	0.74	0.83	0.54
CWJ1	Chien, Wang, Jeng	0.53	0.62	0.40
TY3	Yeh	0.52	0.56	0.41
TY4	Yeh	0.52	0.56	0.41
YSLP1	Yoon, Song, Lee, Park	0.52	0.66	0.39
HCCPH1	Hsueh, Coover, Chen, Popp, Han, Pardo	0.50	0.59	0.39
LYRS1	Liao, YEH, Roebel, Su	0.47	0.54	0.36

There are 3 subtasks for the QbSH task in MIREX 2011 and JSSLP1 which is the name of algorithm we submit won in 2 subtasks of them.

Table 3. Result for QbSH task

Query By Singing/Humming				
SubID	Participants	Task1A MRR	Task1B MRR	Task2 MRR
TY1	Yeh	**0.93**	0.44	8.74
JSSLP1	Jang, PARK, Song, Shin, JANG, Lee, Lee, Seo	0.90	**0.91**	**9.28**
TY2	Yeh	0.88	0.85	8.74

5. REFERENCES

[1] A. P. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in Proc.7th International Symposium Music Information Retrieval, pp.216-221, Victoria, Canada, Oct2006.

[2] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic non-negative matrix factorization for polyphonic pitch transcription," in Proc. IEEE International Conference on Acoustics, Speech and Signal Process., pp.109-112, LasVegas, U.S.A. April 2008.

[3] G. Poliner, D. P. Ellis, A. F. Ehmann, E. Gomez, S. Streich, B. Ong, "Melody Transcription from Music Audio: Approaches and Evaluation," IEEETrans. Audio, Speech and Language Process., Vol.15,No.4,pp.1066-1074,May2007.

[4] J. -S. R. Jang and M. Y. Gao, "A query-by-singing system based on dynamic programming," International Workshop on Intelligent Systems Resolution (the 8th Bellman Continuum), Hsinchu, Taiwan, pp 85-89, Dec. 2000.

[5] J. -S. R. Jang and M. Y. Gao, "A general framework of progressive filtering and its application to query by singing/humming," IEEETrans. Audio, Speech and Language Process., Vol.16,No.2,pp.350-358,Feb., 2008.

[6] Y.Zhu and D.ShaSha, "Warping indexes with envelope transforms for query by humming," In proc. ACM SIGMOD Int. Conf. on Management of Data, pp. 181-192, 2003.

[7] H.M.Yu, W.H. Tsai, and H.M. Wang, "A query-by-singing system for retrieving karaoke music," IEEETrans. one Multimedia, Vol.10,No.8,pp.1626-1637, 2008.

[8] A.Duda, A Nurnberger, and S. Stober,"Towards query by singing/humming on audio databases," Proc. ISMIR,2007.