SINGING MELODY EXTRACTION IN POLYPHONIC MUSIC BY HARMONIC TRACKING

Chuan Cao, Ming Li, Jian Liu and Yonghong Yan
Thinkit Speech Lab., Institute of Acoustics,
Chinese Academy of Sciences,
{ccao,mli,jliu,yyan}@hccl.ioa.ac.cn

ABSTRACT

This paper proposes an effective method for automatic melody extraction in polyphonic music, especially vocal melody songs. The method is based on subharmonic summation spectrum and harmonic structure tracking strategy. Performance of the method is evaluated using the LabROSA database [1]. The pitch extraction accuracy of our method is 82.2% on the whole database, while 79.4% on the vocal part.

1 INTRODUCTION

Melody is widely considered as a concise and representative description of polyphonic music and it can be used in numerous applications such as "Query-by-humming" system, music structure analysis and music classification. However, the automatic melody extraction is recognized to be very tough and remains unsolved up to now.

Yet, amount of remarkable work has been done recently. In 1999, Goto for the first time used a monophonic pitch sequence to represent music melody and achieved transcription from real world music with his famous PreFEst algorithm [5]. Klapuri [1] then proposed a perceptual motivated algorithm in 2005. Poliner and Ellis introduced a novel classification approach using SVM theory for the transcription task [2]. Also, Paiva *et al.* [6] and Dressler [4] proposed methods generally based on spectral peaks picking and post-tracking.

In most methods above, pitch information (pitch candidates, instantaneous frequency (IF) estimations or others) is analyzed frame by frame, and then integrated with temporal/spectral restrictions. However in polyphonic music, especially vocal melody songs, some local frames are inevitably dominated by non-melody intrusions and thus local pitch information is polluted somewhat, or even destroyed sometimes. So integration process based on the polluted information could hardly find the true melody at those local frames.

In this paper, we propose a harmonic tracking method attempting to solve this problem. Briefly speaking, we trace a sound with its harmonic structure in frequency domain. Here harmonic structure mainly refers to the harmonic partials' frequencies and their relative amplitudes. If given target harmonic structure for a specific local frame, we could find the partials from the same sound in adjacent frames, by a tracking strategy. In real applications, target harmonic structure is not known priorly and thus has to be estimated from the mixed signal. We analyze the predominant pitch of the mixture to find stable harmonic structure seeds and then use them to track forward and backward. Also, rather than tracing all the harmonic partials, we use subharmonic-summation (SHS) spectrum as the tracking feature for simplicity, which can be considered as an integrative representation of the whole harmonic family. And a verification procedure is needed to make up the gap between full partial tracking and integrative feature tracking.

[1] The database can be downloaded from the web site of: http://labrosa.ee.columbia.edu/projects/melody/

2 METHOD DESCRIPTION

2.1 Subharmonic Summation Spectrum

Subharmonic-summation algorithm used here is based on Hermes' pitch-determination algorithm [3], concluded as:

$$H(f) = \sum_{n=1}^{N} h_n P(nf) \tag{1}$$

where, $H(f)$ is the subharmonic-summation value of the hypothetic pitch value f, $P(*)$ is the STFT power spectrum and h_n the compression factor (usually $h_n = h^{n-1}$).

2.2 Predominant F_0 Estimation

We estimate the predominant pitches (not necessarily melody) frame by frame with the f_0s that maximize the frame-wise SHS spectrum $H(f_0)$, noted as F_p. With the assumption that singing voice dominates in most frames, which accords well with the reality, we can declare that most of F_p belong to the singing melody. Further processing is generally based on this F_p.

2.3 Stable Harmonic Structure Detection

Here, stable harmonic structure refers to the harmonic structure that dominates the mixture for some time no shorter than θ_s, the stable length threshold. Since pitches from the

same sound have good temporal continuity, we can easily recognize stable harmonic structures by analyzing F_p. A sequence of continuous pitches from F_p longer than θ_s indicates a stable harmonic structure defined above. Notably, we store their time axis start positions in P_e.

2.4 Harmonic Tracking and Identity Verification

As referred above, we use the SHS spectrum to track harmonic structure instead of partials' IF for feasibility and simplicity concerns. For a specific frame, pitch candidates F_{cand} are selected only if they are close enough to the last confirmed pitch and also they should indicate local maxima in the SHS spectrum. Since the locality, these F_0 hypotheses may be false and indicate an invalid pitch value, so a verification procedure follows. We try to use timbre information and calculate the correlation of relative amplitudes between the hypothetic harmonic family and the confirmed harmonic family. If the correlation is larger than the identity threshold, the hypothetic pitch survives in the F_{cand} pool. Then the F_0 hypothesis with the biggest saliency is selected to be confirmed and the tracking process goes on.

Predominant pitches at every P_e are utilized to initialize the process and it goes forward and backward until no F_0 hypothesis survives. All the pitches from every track are considered as a whole and represent the harmonic structure they belong to. Because of the backward and forward mechanism, we do not guarantee that there is only one harmonic structure valid at a specific frame. So a following mapping algorithm is needed.

2.5 Final Pitch Streaming

For every two competing harmonic structures (which have temporal overlapping part), pitches of the overlapping part are decided as follows: 1. Saliency of the two overlapping parts are calculated respectively by summing saliency of all the pitches in that part. 2. The part with higher saliency is reserved and the other is removed.

After all competing pairs have been processed, the final pitch stream is formed.

3 EXPERIMENT RESULT

3.1 Experiment Description

For evaluation, we chose the database released by LabROSA of Columbia University, which was originally made as part of MIREX 2005 Audio Melody Extraction test set. The database is composed of 9 vocal songs and 4 midi music. Extraction results were simply compared to the ground-truth pitch sequences for melody frames, with the tolerance of 1/4 tone. And accuracy of the predominant pitch sequences was also calculated for comparison. As we focus on the singing melody extraction, we also tested our system on a database that contains vocal only songs, 9 vocal songs from the LabROSA database and 4 pop songs from ISMIR2004 Audio Melody Extraction test set.

3.2 Results

As seen in table 1, raw pitch accuracy of the proposed method is 82.23% on the whole LabROSA database, compared to that of the predominant pitches 78.30%. And tests on vocal only songs showed accuracy of 79.39% for the final pitches, while 74.12% for the predominant pitches.

File	A_f (%)	A_p(%)	File	A_f(%)	A_p(%)
track01	84.75	83.43	track11	95.68	91.39
track02	59.47	61.52	track12	99.26	95.98
track03	77.00	68.82	track13	83.19	83.59
track04	73.31	70.29	pop1	78.62	75.44
track05	87.29	83.46	pop2	83.32	79.34
track06	66.84	55.36	pop3	82.65	70.50
track07	80.14	77.05	pop4	88.34	75.51
track08	82.24	80.12	Overall		
track09	86.09	76.43	LabROSA	82.23	78.30
track10	91.35	84.27	Vocal	79.39	74.12

Table 1. Results on the LabROSA database, A_f represents the raw pitch accuracy of the final pitch, while A_p the accuracy of the predominant pitch.

4 CONCLUSION AND FUTURE WORK

The improvement upon predominant pitch is 3.87% on the whole database and 5.27% on the vocal only set. Actually, the improvement could be considered much more significant than the figures shown above since the non-organized predominant pitches are grouped and organized in harmonic structure units, which can be taken as a whole for further considerations. Since the tracking and verification rules used are quite primary, the method can be improved further.

5 REFERENCES

[1] A.Klapuri. "A perceptually motivated multiple-f0 estimation method," In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp291-294, 2005.

[2] G.E.Poliner and D.P.W.Ellis. "A classification approach to melody transcription," In *Proc.6th International Conference on Music Information Retrieval*, pp161-166, 2005.

[3] Dik Hermes. "Measurement of pitch by subharmonic summation," *Journal of Acoustic of Society of America*, vol.83, pp.257-264,1988.

[4] K.Dressler. "Extraction of the melody pitch contour from polyphonic audio," In *Proc.6th International Conference on Music Information Retrieval*, 2005.

[5] M.Goto. "A real-time music scene description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," In *Speech Communication*, vol.43, no.4, pp.311-329,2004.

[6] R.P.Paiva, T.Mendes, and A.Cardoso. "On the detection of melody notes in polyphonic audio," In *Proc.6th International Conference on Music Information Retrieval*, pp.175-182, 2005.