MELODY EXTRACTION FOR MIREX 2016

Juan J. Bosch and Emilia Gómez
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
juan.bosch@upf.edu, emilia.gomez@upf.edu

ABSTRACT

This abstract presents our submission to the MIREX 2016 melody extraction task, whose goal is the identification of the melody pitch sequence from polyphonic musical audio. This approach combines two salience functions, one based on a source-filter model and another one based on harmonic summation. Melody pitch tracking is based on the characterisation of pitch contours, and the selection of melody contours based on a set of heuristic rules.

1. INTRODUCTION

During the PHENICX project [7], we have evaluated melody extraction algorithms in the context of symphonic music. using an annotated collection of symphonic music (ORCHSET). This dataset was annotated with an analysis of agreement between different annotators [3]. The melody in this repertoire is not played by a single instrument, but usually instrument sections which often alternate, and sometimes is not energetically predominant. This poses many challenges to state-of-the-art algorithms, whose accuracy is generally much lower when dealing with such data [3]. Source-filter models achieved the best performing accuracy on this data [6].

In previous MIREX evaluation campaigns, one of the best performing algorithms in terms of overall accuracy is [8]. This approach computes a salience function based on harmonic summation, and then creates and characterises pitch contours for melody tracking. Voicing detection is one of the strong aspects of this method, even though there is a potential room for improvement since timbre information is not exploited. While this approach works specially well in vocal music, results obtained in MedleyDB dataset [1] showed a drop of 19 percentage points when comparing the overall accuracy obtained in vocal vs. instrumental pieces. In more complex scenarios such as symphonic music, pitch estimation accuracy is degraded, partially due to the simple salience function. This method seems not to be able to cope with the high spectral density, and strong accompaniment. The motivation behind the present MIREX submission is to combine the estimation accuracy obtained with a generative approach based on a source/filter model with the benefits of using pitch contour characteristics for pitch tracking. In MIREX 2015, this approach [4] obtained the highest overall accuracy results in comparison to the rest of submitted approaches, in datasets containing orchestral music (ORCHSET), indian music (INDIAN) and also best (ADC04) and second best (MIREX05) results on datasets containing R&B, jazz and pop. Results were found to be more average in the case of karaoke recordings of chinese songs.

2. METHOD

The proposed method combines a salience function based on a source-filter model (SIMM) [5,6] with a method based on pitch contour characterisation [8]. This method is essentially the same as in MIREX 2015 submission [4], with different parameters. The approach is described in detail in [2], where it is also compared with a method based on pitch contour classification. The source code is also available [1] .

In this approach, the spectrogram of a musical audio signal is modelled as the sum of the leading voice and accompaniment. The leading voice is modelled with a source-filter model, and the accompaniment is modelled with a Non-negative Matrix Factorization (NMF). Candidate melody pitch contours are created from the computed salience function, by grouping pitch sequences using auditory streaming cues. Finally, the melody is estimated using contour characteristics and smoothness constraints. We consider a frequency range between 55 Hz and 1760 Hz, and estimate melody pitch values every T=0.01s.

2.1 Pitch Salience Estimation

We model the spectrum X of the signal as the lead instrument plus accompaniment $\hat{X} = \hat{X}_v + \hat{X}_m$. The lead instrument is modelled as: $\hat{X}_v = X_\Phi \circ X_{f_0}$, where X_{f_0} corresponds to the source, X_Φ to the filter, and the symbol $\circ$ denotes the Hadamard product. Both source and filter are decomposed into basis and gains matrices as $X_{f_0} = W_{f_0} H_{f_0}$ and $X_\Phi = W_\Gamma H_\Gamma H_\Phi$ respectively. H_{f_0} corresponds to the pitch activations of the source, and can also be understood as a representation of pitch salience [5]. The accompaniment is modelled as: $\hat{X}_m = \hat{W}_m \hat{H}_m$, leading to Eqn. (1).

$$X \approx \hat{X} = (W_\Gamma H_\Gamma H_\Phi) \circ (W_{f_0} H_{f_0}) + W_m H_m \quad (1)$$

Several parameters need to be specified: the number of bins per semitone (U_{st}), the number of possible elements

[1] https://github.com/juanjobosch/SourceFilterContoursMelody

Algorithm	OA	RPA	RCA
BG1	**0.592**	**0.663**	**0.808**
BG2	0.531	0.616	0.751
IIY1	0.320	0.348	0.670
KON1	0.214	0.208	0.514
WFJY3	0.186	0.267	0.569
WFJY1	0.185	0.260	0.561
WFJY2	0.157	0.222	0.507
FJ1	0.141	0.184	0.451
FJ2	0.099	0.182	0.454

Table 1. Evaluation results for the ORCHSET dataset, ordered by Overall Accuracy. Results provided for: RPA: Raw Pitch accuracy, RCA: Raw Chroma accuracy and OAC: Overall Accuracy. Bold font indicates the maximum value for each metric. The proposed methods are BG1 and BG2

Algorithm	INDIAN	ADC04	MIREX05	MIREX09
BG1	0.732	0.688	0.581	0.462
BG2	0.763	0.697	0.637	0.584
FJ1	0.681	0.603	0.586	**0.765**
FJ2	0.602	0.556	0.474	0.650
IIY1	**0.844**	**0.718**	**0.673**	0.679
KON1	0.831	0.623	0.577	0.700
WFJY1	0.801	0.703	0.573	0.753
WFJY2	0.816	0.667	0.576	**0.766**
WFJY3	0.798	0.698	0.570	0.754

Table 2. Overall accuracy results for 4 different datasets. Bold font indicates the highest values for any of the methods.

of the accompaniment (R), the number of atomic filters in W_Γ (K), and the maximum number of iterations (N_{iter}). Parameter estimation is based on Maximum-Likelihood, with a multiplicative gradient method [6]. In each iteration the parameters are updated in the following order: H_{f_0}, H_Φ, H_M, W_Φ and W_M. The computed salience function is then adapted to a pitch contour creation process, by the combination with a salience function based on harmonic summation, as detailed in [2].

2.2 Pitch Contour Estimation and Melody Selection

From the lead enhanced salience function, we create melody pitch contour candidates by grouping sequences of salience peaks which are continuous in time and pitch, as performed in [8] [2]. Created contours are characterised by a set of features: pitch (mean and deviation), salience (mean, standard deviation), total salience and length. Finally, three further steps are conducted: voicing detection, octave error minimisation (pitch outlier removal), and final melody selection. Previously calculated characteristics are used in this stage to filter out non-melody contours.

3. RESULTS

The evaluation methodology in MIREX compares the sequence of pitches estimated as melody against the ground truth pitch sequence, and focuses on both voicing detection and pitch estimation itself. An algorithm may report an estimated melody pitch even for a frame which is considered unvoiced. This allows the evaluation of voicing and pitch estimation separately. Voicing detection is evaluated using voicing recall and voicing false alarm rates. Pitch estimation is evaluated with Raw Pitch Accuracy (RPA), which is the proportion of melody frames in the ground truth for which the estimation is considered correct (within half a semitone of the ground truth). Raw Chroma Accuracy (RCA) is also a measure of pitch accuracy, in which both estimated and ground truth pitches are mapped into one octave, thus ignoring the commonly found octave errors. Finally, Overall Accuracy (OA) measures the proportion of frames that were correctly labelled in terms of both pitch and voicing. Further details about the metrics can be found in [9]. The evaluation in MIREX 2015 has been performed in five datasets, which contain: north indian vocal classical music (INDIAN), pop, rock, jazz, Rock, R&B, solo classical piano (ADC04, MIREX05), karaoke singing of chinese songs with synthetic accompaniment (MIREX09), and finally our dataset containing symphonic music recordings (ORCHSET) which was used in MIREX for the second time, and is publicly available [3] .

Table 1 presents three metrics results on the symphonic music dataset, and Table 2 shows overall accuracy results on four different datasets in MIREX. Similarly to previous year (2015), our best results in comparison to the rest of the approaches were obtained on the orchestral music dataset. The raw pitch accuracy obtained by our method (BG1) reaches 0.66, which is in relative terms about 86.7% higher than the second best approach (RPA = 0.35). These results show that orchestral music is still very challenging for state of the art melody extraction methods. Table 2 shows that most melody extraction methods are generally tailored for vocal music, since they obtain higher accuracies on datasets containing few or none instrumental melodies. The proposed method BG2 obtains second best results on MIREX05, and very competitive overall accuracy results for ADC04, and even INDIAN, which shows that not only is our method adequate for instrumental music, but also for vocal music from several music genres. Other approaches based on neural networks trained on vocal data are however better able to identify the melody in karaoke recordings of chinese songs with synthetic accompaniment (MIREX09).

The value of the parameters used for pitch contour creation and melody selection plays an important role in the accuracy of the methods for each dataset. Both of our submissions consist of the same algorithm with different parameters. BG2 uses the same set of contour creation and voicing threshold parameters as in [8], while the parameters in BG1 allow the creation of more contours, and has a more relaxed voicing threshold. The results show that BG2 is more convenient for all datasets except for orchestral music. In the case of Orchset, since the dataset is mainly voiced, BG1 achieves a higher RPA and OA.

[2] http://essentia.upf.edu/

[3] http://mtg.upf.edu/download/datasets/orchset

4. ACKNOWLEDGEMENTS

We would like to thank the IMIRSEL team for running MIREX, and the Digital Media Research Center at the Korea Electronics Technology Institute (KETI) team for leading the audio melody extraction task. This work is partially supported by the European Union under the PHENICX project (FP7-ICT-601166) and the Spanish Ministry of Economy and Competitiveness under CASAS project (TIN2015-70816-R) and Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

5. REFERENCES

[1] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello. Medleydb: a multitrack dataset for annotation-intensive mir research. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 155–160, 2014.

[2] J. Bosch, R. M. Bittner, J. Salamon, and E. Gómez. A comparison of melody extraction methods based on source-filter modelling. In *Proc. ISMIR*, New York, Aug. 2016.

[3] J. Bosch and E. Gómez. Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms. In *9th Conference on Interdisciplinary Musicology – CIM14*, Berlin, Dec. 2014.

[4] J Bosch and E Gómez. Melody extraction by means of a source-filter model and pitch contour characterization (mirex 2015). *Music Inform. Retrieval Evaluation eXchange (MIREX)*, 2015.

[5] J. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *Sel. Top. Signal Process. IEEE J.*, 5(6):1180–1191, 2011.

[6] J. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *Audio, Speech, Lang. Process. IEEE Trans.*, 18(3):564–575, 2010.

[7] E. Gómez, M. Grachten, A. Hanjalic, J. Janer, S. Jorda, C. Julia, C. Liem, A. Martorell, M. Schedl, and G. Widmer. Phenicx: Performances as highly enriched and interactive concert experiences. *Open access*, 2013.

[8] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. Audio. Speech. Lang. Processing*, 20(6):1759–1770, 2012.

[9] J. Salamon, E. Gómez, D. Ellis, and G. Richard. Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. *IEEE Signal Process. Mag.*, 31:118–134, 2014.